Running Head: Measurement at the knowledge level

A Theory of the Measurement of Knowledge Content, Access, and Learning

Peter Pirolli
Xerox Palo Alto Research Center

and

Mark Wilson
University of California, Berkeley

Abstract

We develop an approach to the measurement of knowledge content, knowledge access and knowledge learning. This approach has two elements: First we describe a theoretical view of cognition, called the Newell-Dennett framework, which we see as being particularly favourable to the development of a measurement aproach. Then, we describe a class of measurement models, based on Rasch modeling, which we see as being particularly favourable to the development of cognitive theories. Knowledge content and access are viewed as determining the observable actions selected by an agent in order to achieve desired goals in observable situations. To the degree that models within the theory fit the data at hand, one considers measures of observed behavior to be manifestations of intelligent agents having specific classes of knowledge content and varying degrees of access to that knowledge. Although agents, environment, and knowledge are constitutively defined (in terms of one another), successful application of our theory affords separation of parameters associated with the person from those associated with the environment. We present and discuss two examples of measurement models developed within our approach that address the evolution of cognitive skill, strategy choice and application, and developmental changes in mixtures of strategy use.

A Theory of the Measurement of Knowledge Content, Access, and Learning

A defining feature of modern day cognitive psychology is its theoretical admission of mental states and processes. The complexity of observed behavior is assumed to be a manifestation of unobservable mental states and processes interacting with a complex embedding environment. Under a prevalent approach to cognitive psychology, mental states and processes and their resulting behavioral manifestations are shaped by *knowledge*. For a variety of reasons, one could argue that knowledge ought to be the most scientifically interesting aspect of human psychology. Much, if not most, of the behavioral variability of humans is attributed to knowledge differences arising from different enculturation histories. Our everyday folk psychologies are typically couched in terms of knowledge and intention. Efforts aimed at improving education, artifacts, and community life are usually cast in terms of shaping or exploiting knowledge.

In this paper, we propose a framework and theory for measuring knowledge and changes in knowledge. We build upon what we call the Newell-Dennett framework (Dennett, 1988; Newell, 1982) for the observation and explanation of intelligent activity at the *knowledge level*. Informally, knowledge-level theories explain or predict intelligent activity based on the knowledge that an agent may use to achieve its goals in the environment in which it exists. This framework is elaborated with assumptions about quantitative invariants that might be measured about states of knowledge content, mixtures of knowledge content, degree of access to that content, and changes in states, mixtures, and degree of access. For instance, the approach is meant to deal with situations in which an agent acquires one or more strategies for dealing with a task, where degree of access to those strategies changes with experience, and where the particular mixture of strategies used also shifts over time.

We elaborate the Newell-Dennett framework by drawing from the field of psychological measurement. In that area, the manifest behavior of the agent consists of observed responses to questions, problems, or formulaic situations--in other words, what are generically termed "test items" or just "items." The covert aspects of the agent, including strategies and intentions, correspond to the latent structure of those responses, generally characterized through the parameters of a model to be estimated. We use a measurement approach that is founded on the work of Rasch (1960). A feature of Rasch measurement, when it is applicable, is that it affords the separation and quantification of

variables that are assumed to be implicitly and conjointly influencing the observed behavior. This feature is important when observing knowledge-based behavior that is a function of variables associated with the person and variables associated with the embedding environment.

We see the research reviewed here as having two important implications for psychological science. First, we see the combination of broad assumptions about the nature of knowledge-level systems with assumptions about their observation from a sound measurement perspective as an attempt to make the knowledge level into a serious tool for scientific psychology rather than the informal, albeit interesting, level of explanation, which seems to have been its typical use so far. Indeed, most if not all of the published illustrations of knowledge-level explanations are everyday "folk theory" explanations of mundane behavior, and in general, the tendency is to use knowledge level explanations as an informal waystation on the road to mechanistic process explanations (Newell, 1993). We expect that the acknowledged heuristic value of the knowledge level in psychological research can be improved by elaborating it with quantitative measurement. This is because we see quantitative measurement itself as an accelerator of scientific progress, and because the specific Rasch approach we develop is associated with an extensive and well-honed set of inferential tools. Second, we see the flexible measurement methodology presented here as a valuable addition to the armentarium of research psychology, making the advantages of advances in item response modeling available for the analysis of experimental and quasi-experimental designs (Wilson, 1993; Wilson & Adams, 1992). We do not claim, in this paper, that either the cognitive theory approach we describe or the measurement approach we describe are necessarily the only possible choices for such purposes. What we do see is that each has features that makes it a good fit with the other, that this correspondence has some helpful advantages in focusing both theoretical positions, and data analysis, and that such a correspondence is necessary to scientific progress in the field of cognitive science.

We present the elements of our approach to measurement in the context of hypothetical educational situations involving learning to program. Our first example is an application of the theory to a pool of four studies of the development of cognitive skill for programming and is similar to situations commonly found in educational settings. We then present a second model addressing the development of problem-solving strategies for problems in which children are asked how a beam on a fulcrum will balance for different configurations of weights and weight placements. This model will be used to illustrate

how the model deals with developmental changes in the mixture of knowledge content that is used in given situations.

Theoretical Orientation

The Newell-Dennett Framework for Observation of Knowledge Content

Over the course of twenty years, Newell (Moore & Newell, 1973; Newell, 1982; Newell, 1990; Newell, Yost, Laird, Rosenbloom, & Altmann, 1992) developed a set of ideas about system levels that provides a way of understanding how physical systems could be characterized as *knowledge systems*. A very similar set of ideas has been developed by Dennett (1988) in his discussion of *intentional systems*[1], which derives from the work of Brentano (1874/1973). The Newell-Dennett formulations provide a large part of the epistemological framework for our measurement approach.

In the frame of reference developed by Newell and Dennett, observers ascribe knowledge to behaving systems.[2] The *knowledge level* was developed (Newell, 1982) as a way to address questions about the nature of knowledge and the nature of ascribing knowledge to an agent. A key assumption is that knowledge-level systems can be specified completely by reference to their interaction with the external world, without reference to the mechanical means by which the interactions take place. A knowledge-level system consists of an *agent* behaving in an *environment.* The agent consists of a set of *actions*, a set of *perceptual devices*, a *goal,* and a body of *knowledge.* Goals are preference functions over the joint behavior of the agent and environment. The preference functions (goals) are those of the agent. The operation of such systems is governed by the *principle of rationality*: if the agent knows that one of its actions will lead to a situation preferred according to its goal, then it will *intend* the action, which will then be taken if it is possible. The precise nature of the principle of rationality is not discussed at length by Newell (although he was clear in stating that it was not a normative sense of rationality). However, others such as Anderson (1990) and Dennett (1988, ch. 8) assume that the principle derives from the notion of adaptive fit.

In essence, then, the basic observations at the knowledge level are statements of the form:

---

[1]To clarify terminology, what we are calling "knowledge" corresponds to Newell's (e.g., 1982, 1990) use of the term. This, in turn, corresponds to Dennett's use of "belief," which is consistent with common philosophical usage.

[2] Dennett defined an observer who describes a system using an intentional vocabulary (e.g., "know", "believe", "think") as one taking an *intentional stance.*

"In situation *S*, agent *A* behaves as if it has knowledge *K*."

or, as Newell (1982, p. 105) stated, knowledge is "whatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality."

Example: Knowledge-level Observation of Knowledge Content

To illustrate the notion of knowledge-level observation–the ascription of knowledge content to a behaving agent–we consider the general situation that provided data for our first model, discussed in detail below. The data set for this first model application came from four experiments (Bielaczyc, Pirolli, & Brown, 1995; Pirolli & Recker, 1994; Recker & Pirolli, 1994) , in which data were collected by an intelligent tutoring system (ITS), called the CMU Lisp Tutor (Anderson, Boyle, Corbett, & Lewis, 1990).

These studies were conducted under experimental conditions, and although the instructional technology was uncommon, the general pedagogical situation and knowledge-assessment problems were quite familiar. Students had studied or listened to some lesson materials and then solved some exercise problems. From their exercise solutions, we wanted to infer what they knew and how well they knew it. To which materials had they attended? What prior skills did they seem to have already mastered? What skills had they failed to learn? How fluently could they use their knowledge?

Suppose we observed students working on their program-writing exercises and wanted to infer the instructional examples to which they had attended. From patterns of problem solving behavior, we wanted to make statements about whether or not a student "knew" elements from one or more of the examples used in their instructional materials and the degree to which they "knew" their programming skills. For instance, suppose that the observed students may or may not have studied either of the example Lisp programs presented at the top of Figure 1 and among their exercises they must write the two Lisp programs presented at the bottom of Figure 1.

Our basic expectations of how knowledge will transfer from the examples to the exercise problems are summarized in Table 1. Our research hypothesis was that, due to similarity in problem structure in Lisp (Pirolli & Recker, 1994), that students who know the *Sumall* example will show higher transfer of example knowledge to their solution of the *Fact* exercise than for the *Length* exercise, and students who know the *Carlist* example should show the opposite pattern, with *Length* showing higher tranfer than *Fact*. Reasoning backwards from observations of the pattern of problem solving on the *Fact* and

Length exercises, we should be able to make inferences about the degree to which students know the *Sumall* and *Carlist* examples.

Insert Figure 1 and Table 1 about here

Suppose, that a student of Lisp is in the midst of writing her first recursive program *Fact*. Consider the situation-action sequence schema in Figure 2. Imagine that the student has completed the code on the left of the arrow in Figure 2. We then observe that she then completes the program with the specific action of writing the underlined code to the right of the arrow. The coded element, (- *n* 1), subtracts 1 from the input argument *n* and it is the appropriate *recursive step* to take in many simple recursive programs involving numeric inputs. An appropriate knowledge-level statement by an observer of the student's action in this situation might be something like:

> In her first encountered situation involving the goal of coding a recursive step on a numeric input, the student behaved as if they knew the analogous recursive step of the *Sumall* example

This statement captures an observation of some element of knowledge. That is, we have observed that the student is in a particular *state of knowledge content.*

Insert Figure 2 about here

Knowledge Level vs Symbol Level

We also draw upon Newell's (1982; 1990) distinction between *knowledge level systems* and their mechanistic information-processing descriptions as *symbol level systems*. This distinction leads us, in the next section, to discussion of two parallel distinctions of relevance. The first is the distinction between *knowledge content*, which we associate with the knowledge level, and *knowledge access*, which is carried out by symbol-level mechanisms. The second is the distinction between *knowledge-level learning* (changes in knowledge content) and *symbol-level learning* (changes in knowledge access).

Newell (1982; 1990) developed the notion of system levels as a proper approach to the explanation of cognition. The symbol level is defined by a medium of formal patterns that yield behavior through mechanistic computation. Many explanations and models in cognitive psychology are cast at this level, especially those cast as computational programs specified in *cognitive architectures* (Anderson, 1983; Anderson, 1990; Newell, 1990). It is generally assumed that symbol-level systems are realized by lower-level systems that ultimately are grounded in physical media and laws.

In the Newell-Dennett approach, knowledge at the knowledge-level is defined functionally (Newell, 1982). Knowledge is defined as a function mapping situations and intentions onto behavior. Knowledge does not reside in any particular state-like structure defined at the symbol level, although symbol-level structures may be involved in the computation of the knowledge function, just as structures or states in a computer are part of the computation of input-output functions. This functional definition of knowledge is adopted by Newell and Dennett because knowledge about the world cannot be captured in extension by a finite physical structure–for instance, as a structure containing or listing each element of knowledge. Newell clearly states that knowledge is defined "in terms of what it does...knowledge, though a medium, is embodied in no medium-like passive physical structure...[with a] state-like physical structure" (Newell, 1982, p. 105). The potential set of things that could be known about the world, and more technically, which could be ascribed to a potential agent, is unbounded:

> What the computational system generates are selections of actions for goals, conditioned on states of the world. Each such basic means-ends relation may be taken as an *element* of knowledge. To have the knowledge available in extension would be to have all these possible knowledge elements for all the goals, actions and states of the world discriminable to the agent at the given moment. The knowledge could then be thought of as a giant table full of these knowledge elements, but it would have to be an infinite table. Consequently, this knowledge (ie., these elements) can only be created dynamically in time. If generated by some simple procedure, only relatively uninteresting knowledge can be found. Interesting knowledge requires generating only what is relevant to the task at hand, ie., generating intelligently. (Newell, 1982, p.108, italics in original)

Thinking about knowledge as a dynamically allocated table of elements is a useful formulation that we will return to in a moment. These elements may also be thought of as elements of an infinite set defined by a relation over goals, perceptual states, and actions.

Examples of symbol-level models would be the production system models of Lisp programming used in work related to our first example (Anderson, 1984; Anderson, Conrad, & Corbett, 1989; Anderson, Pirolli, & Farrell, 1988; Pirolli, 1985; Pirolli & Anderson, 1985). Individual production rules and proposition-like elements in production-system working memory are elements of symbol-level analysis. The production rules formally represent elements of cognitive skill and the proposition-like elements formalize facts. The production system architectures that run these models employ specific mechanisms for selecting and applying production rules in given situations, based on ACT* (Anderson, 1983). These mechanisms embody such principles as ordering production rules in given situations based on their strength and situation-specificity. As we discuss next, symbol-level mechanisms, such as these, determine the context-dependent access of knowledge available to the system.

We must emphasize that general production systems are just a formal notation and, depending on one's theoretical stance, they may be used for symbol-level models or knowledge-level models. Some specific production system architectures such as ACT* (Anderson, 1983) and Soar (Newell, 1990), whose mechanisms are integral to the prediction of cognitive states and processes, are symbol-level models. The production system models used in our intelligent tutoring example below are, we claim, knowledge-level models, because they formalize what knowledge has been exhibited without empirical claim about the cognitive processes by which the knowledge came to be exhibited. In a way, a knowledge-level production rule analysis of behavior is like a formal grammatical analysis of an utterance: It describes the underlying deep knowledge that was exhibited without commitment to the manner in which the knowledge came to be expressed.

Knowledge Access is Defined by the Symbol Level

In general outline (Newell, 1982), a *representation scheme* is defined at the symbol level as a combination of data structures and processes specified in some architecture. An architecture plus knowledge representation scheme determines a context-dependent *knowledge access function* for the knowledge system. Consider the infinite table of knowledge mentioned in the Newell quote above. For any given environmental context, the symbol-level representation scheme predicts measurable properties about the realization of that knowledge in that context. Some elements of knowledge will be more accessible than others depending on context. One may think of the the symbol-level representation scheme as defining a degree-of-access function over the table of knowledge content available to the system.

The term "access" need not be limited to mean the retrieval and interpretation of knowledge structures internal to a person. Depending on how we construe our knowledge-level analysis, we might consider a complete system of a person plus external media as the knowledge-level system, and access to mean the interpretation of those external media as well as the retrieval of internal symbol patterns. Mathematically, in our proposed theory, degree-of-access is assumed to be a conjoint function of the environment and the person. For instance, in our first example, we deal with people learning to program. One aspect of the developed model addresses the observation that a person with a relevant programming example in their environment behaves more knowledgeably than one without the example. That is, the person has greater access to the knowledge or, to state things differently, the environment-with-relevant-example has a greater affordance for that knowledge-based behavior than the environment-with-irrelevant-example.[3]

Measurement Spanning Knowledge and Symbol Levels

The Newell-Dennett notions of system levels and the assumption of an observer ascribing knowledge to a system behaving in context provide us with a basic observational framework. We will cast our measurement theory as spanning the knowledge and symbol levels. We will be concerned with the measurement of properties of knowledge access functions, without detailed concern with the specific computational, symbol-level mechanisms that give rise to those properties. For instance, the measurement model we develop for Lisp programming in our first example aims to measure degree of knowledge access (i.e., the proficiency with which people exhibit specific Lisp programming skills), but the measurement model is not concerned with how a specific production system model might produce such measurements. In this sense, our measurement theory sits at the interface between the knowledge level and symbol level. The measurement theory is not a mechanistic theory at the symbol level but it addresses measurable properties associated with the knowledge access function that ultimately rely on mechanisms at the symbol level. This notion is elaborated by considering issues related to the assessment of learning.

Learning as Change in Knowledge Content and Knowledge Access

The analysis of learning in the context of Newell's system levels has generated considerable discussion in recent years (Anderson, 1989; Dietterich, 1986), and there is

[3]The general observation is that subjects use examples by interpreting their external form, rather than first memorizing them and then working from memory (Chi, Bassok, & Lewis, 1989; Pirolli & Anderson, 1985).

no agreed-upon viewpoint (Agre, 1993). For our purposes, we adopt some broad characterizations of learning consistent with a broad class of theories:

1. At the knowledge level, knowledge content grows monotonically and in discrete quanta through interaction with the environment. That is, we assume that knowledge is only added to the system.[4] Changes at the knowledge level are also (necessarily) changes at the symbol level because the knowledge content of the knowledge level must be implemented in the symbol level. (Note that we view forgetting as a symbol-level event, so that this does not rule out memory failure, etc.)

2. There are additional changes at the symbol level that are reflected by changes in the access function to existing knowledge—that is, without changing the knowledge content of the system. Improvements due to repeated practice and forgetting due to disuse are commonly interpreted as effects of symbol-level learning.

We will call changes in knowledge content *knowledge-level learning*, whereas changes in properties of knowledge access are called *symbol-level learning*. Although this terminology differs from that used in the machine learning literature (Dietterich, 1986), we believe it is consistent with Newell's (1982) original formulation.

Newell and Dennett's formulation of the knowledge level is concerned with response functions (mappings of environmental situations onto behavioral responses) with no concern for mechanism. Knowledge content, perceptions, actions, and goals, situated in an environment, define response functions, R, that map onto behavior.[5] Our knowledge-level observations concern the response function R. The system knows to do something, but the knowledge level does not specify how it is done. The symbol level specifies those mechanisms. A property of those mechanisms is the propensity with which knowledge-level responses occur, where propensity might be operationalized, for instance, as response speed or response probability. Knowledge-access properties are properties of R, or $P(R)$. Changes in R necessarily mean changes at the symbol level, since the knowledge level is mechanistically realized by the symbol level. But, changes in $P(R)$ may take place without changes in R and so, strictly speaking, may not be observable from the pure knowledge-level stance. What we are proposing is an

[4]For instance, incorrect beliefs can not be deleted from the system, but may be blocked from manifesting themselves by more correct beliefs. Other schemes for the development of knowledge are conceivable (see, Flavell, 1972). This particular assumption is also consistent with the arguments of Anderson (1989)

[5] A person may intend an *action* such as "intend to pick up a coffee mug" whereas the *behavior* carrying out the action is the actual physical manifestion carried out, which may vary every time the action is carried out.

embellishment of knowledge-level characterization with a minimal, nonmechanistic, quantitative characterization of the symbol-level as embodied in measurable access properties $P(R)$. This is our motivation for distinguishing symbol-level learning from knowledge-level learning.

Example: Intelligent Tutoring Systems as Knowledge-Level Recording Instruments

The ITS used in our studies observes students as they write their programs, and makes inferences about their state of knowledge by making use of a formal production system model of program-writing skills. To a large extent, this approach is motivated by studies (Anderson, 1993; Pirolli, 1991; Polson, Bovair, & Kieras, 1987; Singley & Anderson, 1989) in which it has been productive to assume a correspondence between formal production rules and elements of *procedural knowledge* or cognitive skill. Production rules are formal rule patterns of the form C→A, in which C specifies a condition pattern to which the rule applies and A specifies an action pattern to perform if the rule is executed. In the ITS, the production system models are used to simulate idealized student cognitive skills to solve program-writing problems. The *ideal models* are compared against the input of an actual subject (Anderson, et al., 1990). Context-specific feedback is provided to subjects if they commit programming errors. The Lisp Tutor records external observable situations and a subject's actions and matches these against production rules representing knowledge elements. In this case, the producton system models used by the Lisp Tutor are knowledge-level models. The production rules formalize Lisp knowledge, but no theoretical claim is made for the computational mechanisms that select and execute those rules in the running Lisp Tutor.

Notably, this means that the ITS is a knowledge-level recording instrument. The ITS has an internal table of productions that captures the elements of knowledge that are possible both within and across students. It ascribes these knowledge elements to an observed student when the student exhibits the appropriate behavior. The state of knowledge ascribed to a subject is an overlay on this table of productions. This implements the basic schema for knowledge-level observations of the form, "in situation S, agent A behaves as if it has knowledge K." The ITS knows the mapping of situation to action implied by knowledge elements—represented by the production rules—and mechanically fulfills Newell's (1982) role of observer. Thus, in practice, such an ITS can be viewed as an automated knowledge-ascribing instrument that treats a subject as a knowledge-level system.

As we will illustrate in detail later, we may measure a number of properties of individual elements of programming knowledge. The CMU Lisp Tutor can be used to track the history of each individual cognitive skill represented by its production rules, across the problem situations in which those skills are evoked. That is, we will have a sequence of trials, for each cognitive skill, extracted from the full protocol of behavior exhibited by a student. During the course of observations over several programming problems, a student may be observed to change their basic programming strategy–a change in knowledge content–or to improve in their proficiency in exhibiting a specific cognitive skill–a change in knowledge access–or both. This leads us to consideration of issues concerning the measurement of learning.

Representation and Quantification

We note that, for the purposes of measurement, the semantics of a formal representation of knowledge have to be sufficient for the scientific question at hand. We take this as a pragmatic issue. Like other formalizations in science, the interpretation of formal statements about knowledge rely on the shared understanding of scientific practitioners. As a consequence of this stance, the main issue we are concerned with is whether a formalization of knowledge and its scientific interpretation are sufficient and appropriate for the data at hand. This stance is consistent with the writings of Newell (e.g., Newell, 1982) and Dennett (e.g., Dennett, 1988). For instance, Newell's basic assertion about formal representation of the knowledge level was that "to ascribe to an agent the [symbol] structure *S* is to ascribe whatever the observer can know from [symbol] structure *S*" (Newell, 1982, p. 112). In the example above, we happened to assume a production rule representation of knowledge, but this is not a necessary component of the basic approach.

In addition to representing knowledge content, we will want to measure properties concerning its access. Cognitive psychology has largely ignored measurement issues (Cliff, 1992). However, outside of cognitive psychology, the pursuit of more direct quantitative measurement of theoretical variables has a long history in psychology, and this has principally been applied to knowledge access variables such as reaction times, and the items in IQ tests. This is precisely because quantification is such a strong tool for, and marker of, understanding and control of over the phenomena of interest:

> Quantification in science is inseparable from the experimental method. The hypothesis that a particular variable is quantitative is a substantive hypothesis. It requires that the values of that variable manifest a definite kind of structure. With

> the possible exception of some quantitative measurements open to extensive measurement, evidence for or against the quantitative measurement must be gained by experiment. The gathering of such evidence requires a high degree of experimental control and, often, sophisticated apparatus and methods of observation. It is no accident that the extension of quantification in physics from geometry and statics to dynamics, thermodynamics, and electrical phenomena went hand in hand with experimental techniques and apparatus. Psychological measurement, if it is to be, requires the same kind of advances.(Michell, 1990, p. 86)

In the natural sciences, it is not uncommon to begin textbooks with a chapter on measurement (e.g., Halliday & Resnick, 1970), and this is mimicked by both editions of *Steven's Handbook of Experimental Psychology* (Atkinson, Herrnstein, Lindzey, & Luce, 1988; Stevens, 1951). As we will outline below, our theory derives from *specifically objective measurement* (Rasch, 1960) which (in its determinate version, Fischer, 1973; Glas, 1989; Wilson & Pirolli, 1995) meets the conditions of *fundamental measurement* theory (Krantz, 1964; Luce & Tukey, 1964) whose development was, in part, aimed at placing psychological measurement on the same firm ground as measurement in sciences like physics.

Constitutive Definitions and the Separation of Parameters

As mentioned above, the Newell-Dennett framewwork includes as principaal elements both agents and environments. Both agent variables and environment variables are reflected in the same observation of behavior. This raises concerns about the separation and quantitative representation of agent and environment parameters, given that they must therefore be *constitutively defined* (in terms of one another). The approach we use derives from the measurement work of Rasch (1960).

The marks of quantity are established by ordinal relations and additive structure among the variables of interest (Campbell, 1928; Krantz, 1964; Luce & Tukey, 1964). In the case of extensive properties, such as length, specific ordinal relations are clearly manifest in comparisons of objects of different lengths and additivity is manifest in the manner in which lengths can be concatenated to produce new lengths. For example, with respect to the extensive property "length", we note that two particular objects can be compared and ordered ("longer than") and they can be concatenated in a particular way to yield a new length that can be tested against other lengths. These ordering, comparison, and concatenation operations must meet certain conditions in order to be quantified in a coherent and meaningful way (Michell, 1990).

The problem we have here does not involve extensive variables. Rather, the agent and environment parameters must be constitutively defined, and this raises special issues regarding the separation of the two kinds of variables. The approach to this problem is familiar to anyone who has used the additive factors logic of experimental design.[6] Suppose we were investigating the relation of force, mass, and acceleration in classical physics (which we now characterize as $f = ma$). Suppose we impart forces $f_1, f_2, \ldots f_n$ on masses $m_1, m_2, \ldots, m_m$ and measure accelerations $a_{11}, a_{12}, \ldots, a_{nm}$. This means that a measure of a_{ij} is a measure of the ordered pair $\langle f_i , m_j \rangle$. One can imagine laying out these measures as a two-way table, with one class (e.g., force) along the row headings, a second class (e.g., mass) along column headings, and the third property (measured in conjunction with the other two; e.g., acceleration) in the cell entries in the table. Such a table is illustrated schematically in Figure 3, where the bar lengths on the left of the figure represent the original observations. If variables defined over the two classes of entities and the resultant response variable can be simultaneously scaled so that an ordinal additive (noninteractive) structure results, then one can separate the scales associated with the variables. If we stay in the original metric, there is no apparent way to add forces (f_i) and masses (m_j) to get the accelerations (a_{ij}) , but by performing simultaneous logarithmic scaling on the raw variables such that $\phi_i = \log(f_i)$, $\mu_{j\cdot} = -\log(m_j)$, and $\alpha_{ij} = \log(a_{ij})$ one achieves an additive structure such that $\alpha_{ij} = \phi_i + \mu_j$. Taking two measurements α_{11} and α_{12}, one can measure masses μ_1 and μ_2 independently of the force chosen as $\phi_{\cdot}$, since

$$\alpha_{11} - \alpha_{12} = (\mu_1 - \mu_2) + (\phi_1 - \phi_1) = \mu_1 - \mu_2.$$

Similarly, accelerations α_{11} and α_{21} can be used to compare ϕ_1 and ϕ_2 by eliminating μ_1. Units for the scales can be achieved by appropriately selecting some standard α_{jk} as the zero point and some other $\alpha_{jk'}$ or $\alpha_{j'k}$ as the unit position. This scaling and separation is illustrated on the right side of Figure 3, where the bar lengths correpond to transformed scores, and additivity in represented by the concatenation of bar lengths.

Insert Figure 3 about here

The physics example suggests how separate quantitative scales can be achieved if one can find a simultaneous transformation that maintains the order relation and reveals additivity among observations. Such measurement is addressed formally by *additive conjoint measurement* theory, a kind of fundamental measurement theory (Krantz, 1964;

[6]The following example is based on Rasch (1960) and Andrich (1988).

Luce & Tukey, 1964), which addresses issues not found with the measurement of extensive properties, such as physical length, concerning the separation of scales. Additive conjoint measurement establishes axioms that must be met to satisfy the appropriate order and algebraic structure to quantify constitutively defined variables. Unfortunately, additive conjoint measurement supposes a deterministic framework. This is not consistent with the probabilistic conception of observations at the the knowledge level, as described above. What we need is a probabilistic version of the approach we have portrayed in the physics example.

The aim is to map observations of *manifest* situations and behavior onto measures associated with the *latent* or unobservable knowledge. We think of this mapping as probabilistic rather than deterministic for two possible reasons. First, there is standard problem of an observer's uncertainty about the inferences made from a finite set of observations–in this case about an agent operating in their environment. Second, there is the possibility that knowledge maps onto behavior in a stochastic manner, much as assumed in the competence-performance distinction (Chomsky, 1965). That is, we may want to think of knowledge as response functions which characterize the *probabilties* of behavior in given situations, rather than the exact unique behavior. Either or both of these assumptions lead us to adopt the stance that there is a probabilistic relationship between the latent knowledge-level constructs and manifest observables. This limits the usefulness of approaches such as fundamental measurement (Krantz, 1964; Luce & Tukey, 1964) which are based on at least the logical possibility of deterministic ascriptions.

Example: Separation of Parameters in the Rasch Approach

This brings us to the work of Rasch (1960). To illustrate the essence of the Rasch approach, suppose that an observer dichotomously scores responses made by an agent n in situation i such that $X_{ni} = 1$ means that "subject n acted as if they had knowledge appropriate for situation i " (a "successful" response), otherwise $X_{ni}= 0$ (an "unsuccessful" response). This might occur, for instance, if the observer was scoring performance by students writing their first recursive functions in Lisp, and specific program-writing responses, such as "(- 1 n)", in specific situations, such as those involving the goal to code a recursive step. The Rasch (or logistic) model would then characterize the response probabilities as

$$\Pr(X_{ni} = 1 | \theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \quad (1)$$

$$\Pr(X_{ni}=0|\theta_n,\delta_i)=\frac{1}{1+\exp(\theta_n-\delta_i)}, \quad (2)$$

where θ_n is a parameter characterizing agent n (usually called "ability" in the Rasch literature) and δ is a parameter characterizing situation i (usually called "difficulty" in the Rasch literature).

The response odds are then:

$$\frac{\Pr(X_{ni}=1|\theta_n,\delta_i)}{\Pr(X_{ni}=0|\theta_n,\delta_i)}=\exp(\theta_n-\delta_i). \quad (3)$$

It should be noted that simultaneous transformation of both sides of Equation 3 yields the additive structure

$$\log\left[\frac{\Pr(X_{ni}=1|\theta_n,\delta_i)}{\Pr(X_{ni}=0|\theta_n,\delta_i)}\right]=\theta_n-\delta_i. \quad (4)$$

That is, the log odds of successful response is just a sum of agent and situation parameters.[7]

So, Equation 4 achieves the desirable properties of the classical physics example. The variable associated with the person is separable from that of the external problem, and these variables may be scaled quantitatively. It is noteworthy that the variable separation and scaling afforded by the Rasch approach is not achieved in such a direct fashion by other psychometric approaches (see for example, Carroll, 1988).

Rasch (1960) called this key concept on which he based his models, *specific objectivity*. In his formulation, it is the equivalent of an additive conjoint structure under a probabilistic formulation. Specific objectivity means that a response measure is a conjoint measure of two entities (such as an agent and situation) whose measures can be separated and quantified, similar to the manner discussed above. Rasch stated that specific objectivity holds when

> the result of any comparison of two [agents] ... *is independent of everything else within the frame of reference other than the two* [agents] *which are to be compared* ((Rasch, 1977), p.77, italics in original).

[7] The values of θ_n and δ_i are usually reported as logit scores or logits. A logit is that distance on the knowledge measure that corresponds to odds of success (compared to failure) equal to e, the base of the natural logarithms--approximately 2.7:1.

That is, the parameter describing the agent must be inferentially separable from the parameters describing the environment. This must hold, in a dual fashion, for comparisons of environment parameters also. Rasch (1960) showed that, under mild assumptions, his Rasch model is both necessary and sufficient for specific objectivity (This proof was formalized and extended by Andersen [1977]). It is possible to demonstrate a formal relationship between the Rasch model and additive conjoint measurement, and we do so in another recent paper (Wilson & Pirolli, 1995).

Rasch (1960) established that sufficient statistics (Kendall & Stuart, 1969) could be obtained quite simply for the parameters in his model. Estimators based on such statistics fulfill the dual role of establishing the empirical conditions under which a model applies, and providing the underpinnings for statistical estimation and inference. The argument outlined above, basing the Rasch model on the logic of comparison, has been extended by Masters and Wright (1984) to a family of Rasch models suitable to environment conditions more complex than the dichotomous one (e.g., polytomous categories of response). There is an extensive literature elaborating and refining the estimation and calibration technologies associated with Rasch models (Fischer & Molenaar, 1995; Glas, 1989; Gustafson, 1980; Wright & Douglas, 1977a; Wright & Douglas, 1977b).

Coupling Measures of Knowledge-Level Content and Symbol-Level Access

We now turn to several basic assumptions of the approach, concerning the nature of variables measuring the degree to which a person is in a knowledge-content state and their degree of acess to elements of knowledge in that state, as well as the mapping of manifest observations onto those variables. In our measurement framework, both knowledge-level content and symbol-level access influence behavior. It seems to us that a reasonable first approach to this problem is to think of symbol level learning--changes in the knowledge access function--as being represented by continuous variables along which agents and specific situations are arrayed. Such a view is consistent with common approaches in classical test theory or item response models (e.g., Lord & Novick, 1968). The joint history of an agent with environmental situations over time may alter how agents and environments are arrayed along a knowledge access variable. Such an approach might be used, for instance, in assessing changes in a student's Lisp programming skill over a series of exercise problems. Such a view is also consistent with theoretical approaches common in memory and learning research, where variables such as strength and activation are assumed to represent the accessibility of knowledge, and to change with

specific histories of experience with specific situations. These views are consistent with the idea of symbol level learning as fundamentally incremental.

In contrast, we propose that knowledge level learning is characterized by discrete changes in the knowledge content that forms the basis for actions (i.e., changes in Newell's "knowledge table"). This can be represented by different classes of agents defined as having equivalent states of knowledge content. Such an approach might be used, for instance, in determining which students know one subset of instructional material (such as a particular example) versus another when observed while solving a set of programming exercises. This is the approach we will take in our first example. Such a view resembles the basis for a latent class approach to psychological measurement (Lazersfeld & Henry, 1968). Equivalently, one might also view a specific agent at different points in their history as being in different states of knowledge content, such as having different cognitive strategies or being in different states of domain expertise. This is an approach appropriate, for instance, in examining developmental changes in strategy use. This is the way we will develop our second example, which addresses how children solve balance beam problems.

Now, knowledge content and knowledge access operate simultaneously. Thus, the two types of models, continuous for knowledge access and discontinuous for knowledge content, must also operate simultaneously. Within a given agent-class, there would always be symbol-level learning going on, so the continuous variables would operate *within* classes. In response to this formulation, the approach we outline in this paper describes a combination of latent class and latent variable approaches. It has its statistical roots in the topic of mixture models (Titterington, Smith, & Makov, 1985). A general outline of the approach has been provided by Mislevy and Verhelst (1990), and recent work along these lines has been described by Ekstrand and Wilson (1990), Kelderman and Macready (1990), Mislevy, Wingersky, Irvine, and Dann (1991), Rost (1990), Mislevy and Wilson (1996), and Wilson and Draney (1995).

Probabilistic Knowledge-Content States

To recap, we will assume that measurement models ascribe a probabilistic relationship between unobservable knowledge and observable behavior. We will assume that there are discretely different states of knowledge content. For instance we may identify different discrete strategies for solving a problem, or different discrete experiences with instructional material for a topic. Assuming there are K possible knowledge classes, such as K different problem solving strategies, or K different sets of instructional

backgrounds, then we will find it convenient to represent the knowledge state of person n using a K-dimensional vector

$$\phi_n = (\phi_{n1}, \phi_{n2}, \ldots, \phi_{nK})' \quad (5)$$

of zeros and a single one. That is, $\phi_{nk} = 1$ represents person n being in knowledge-content state k (and consequently, $\phi_{nh} = 0$ for all other states h). Having observed person n interacting with an environment (which may be formalized as responses to "items"), based on that data, we will want to estimate the pattern of probablities across the classes:

$$\hat{\phi}_n = (\hat{\phi}_{n1}, \hat{\phi}_{n2}, \ldots, \hat{\phi}_{nK})' \quad (6)$$

where each $\hat{\phi}_{nk}$ $(0 \leq \hat{\phi}_{nk} \leq 1)$represents the probability that the person is in knowledge-class k. We can also hypothesize a population level parameter π that indicates the probablity of a random member of the population being in each class:

$$\pi = (\pi_1, \pi_2, \ldots, \pi_K)' \quad (7)$$

One way to characterize our approach is by contrasting it with models of *mastery learning* (Atkinson & Paulson, 1972). In such models, one assumes that there are some fixed elements of knowledge whose learning state is characterized by one or more probability parameters. For instance, Corbett, Anderson and O'Brien (1995) use such a model in which they record the probability that a cognitive skill is in a learned (or unlearned) state. The overall state of learning of knowledge content in a mastery learning model is just the instantaneous state of all the variables of all the elements. That is, one may think of the state of mastery as a table of knowledge elements that records the state of variables for each knowledge element. Having multiple knowledge-content classes, $\underset{\sim}{\phi}$, permits the identification of meaningful patterns over the knowledge-element variables that might be associated with different kinds of experiential histories (e.g., different courses of instruction), different kinds of strategies, and so on.

Example: Knowledge-Content Differences Arising from Instruction

Imagine that the cognitive skills for writing recursive programs are represented by a table of knowledge elements and continous variables for each indicating their state of symbol-level learning. We might expect that the learning of these knowledge elements will exhibit different patterns that depend on instructional experience. If we constructed a table of knowledge elements for the cognitive skills involved in programming the recursive functions *Fact* and *Length* in Table 1, we would expect the learning of some knowledge-element variables to be boosted by the student attending to instruction that included the

Sumall example, and other knowledge-element variables to be boosted by attention to instruction that used the *Carlist* example. From observed patterns over the knowledge elements, one could make inferences about whether a student attended to one, the other, both, or neither of the examples.[8]

Example: Transition Paths Through Knowledge-Content States

In developmental psychology, questions may arise concerning the difficulty of transitioning from one state of cognition to another. For instance, we will be concerned with the knowledge-level learning transitions among problem-solving strategies for the balance beam. In educational situations, one may be concerned with modelling the difficulties of different paths through instructional experiences. The use of knowledge-content classes and the notion of knowledge-level learning can be used to address these issues.

Kessler and Anderson (1985) studied a situation in which different orderings of problem-solving experiences in program-writing had different learning effects. Some of their participants first wrote a block of iterative programs followed by a block of recursive programs (iteration-recursion), others had the opposite sequence (recursion-iteration), and others had two blocks of iteration (iteration-iteration), or two blocks of recursion (recursion-recursion). Participants worked on each of the eight programs in each block until the programs ran correctly (i.e. they learned to criterion). Interestingly, the transition from learning recursion to learning iteration was more difficult (in terms of time to criterion) than the transition from iteration to recursion. Kessler and Anderson argued that this was because people think recursive programs are performing iteration if they have not previously seen iterative programs.

One way to think schematically of the iteration-recursion and recursion-iteration transitions through knowledge-content states is depicted in Figure 4. For this situation, one could develop a table of knowledge elements representing the cognitive skills for writing the iterative and recursive programs studied by Kessler and Anderson (1985).[9] In Figure 4, I+ indicates that the knowledge-elements for iteration show criterion-level learning and I- indicates less than criterion-level learning, and R+ and R- have analogous meanings for the recursion knowledge elements. The transition among knowledge-content states for the iteration-recursion learners goes from [I-, R-] to [I+, R-] to [I+, R+],

---

[8] Note that in our model application, below, we experimentally manipulated these instructional experiences rather than infering them from observation.

[9] A production rule model of the skills for the recursive programs of Kessler and Anderson (1985) is presented in Pirolli (1991)

whereas the recursion-iteration learners transition [I-, R-] to [I-, R+] to [I+, R+]. It is this last transition, indicated by the asterisk in Figure 4, that is particularly difficult. Thus, in the notation above, we might construct the knowledge content classes as follows:

$\phi_{n1} = 1$ when person n "knows" neither I nor R--[I-, R-],

$\phi_{n2} = 1$ when person n "knows" I but not R--[I+, R-],

$\phi_{n3} = 1$ when person n "knows" R but not I--[I-, R+],

$\phi_{n4} = 1$ when person n "knows" both I and R--[I+, R+].

Insert Figure 4 about here

Suppose now that we have developed a set of items that are related to I and R: In particular, we believe that certain items require I, certain R, certain both, and certain neither. Within each knowledge content class we will hypothesize that the items can be modeled in a way analagous to the Rasch approach outlined above:

$$Pr(X_{ni} = 1 / \theta_n, \delta_i, \phi_{nk} = 1) = \frac{exp(\theta_n - \delta_{ki})}{1 + exp(\theta_n - \delta_{ki})}, \quad (8)$$

where δ_{ki} is the difficulty of item i within class k, and the other symbols are as defined above. Note that, analagous to its interpretation under the Rasch model, θ_n is aimed at capturing the knowledge access of person n. What is different is the parameter ϕ_n, which is aimed at capturing the the knowledge-content class that they are in. If indeed, we knew, or could assume we knew, each ϕ_n, then we could proceed in a relatively straightforward way to find estimates of θ_n and δ_{ki} within each class. Unfortunately, we do not generally know ϕ_n beforehand (but see later for an example where we do indeed make that assumption), so we need to estimate both simultaneously. In the formulation we adopt

here, we find it useful to bring one more piece of information into play: we assume that the items have been designed to be indicative of certain knowledge classes. This we express through certain relationships among the item parameters δ_{ki}. For example, one might posit, in the context of the Kessler and Anderson situation, that the following holds:

$$\delta_{ki} = \begin{cases} \delta_i, \text{if item } i \text{ requires neither I nor R,} \\ \delta_i + \tau_I, \text{ if item } i \text{ requires I,} \\ \delta_i + \tau_R, \text{if item } i \text{ requires R,} \\ \delta_i + \tau_{RI}, \text{if item } i \text{ requires both I and R.} \end{cases} \quad (9)$$

Then equation 8 becomes a set of equations, depending on which knowledge-content class is applicable:

$$Pr(X_{ni} = 1 / \theta_n, \delta, \phi_{n2} = 1) = \frac{exp(\theta_n - \delta_i - \tau_I)}{1 + exp(\theta_n - \delta_i - \tau_I)}, \text{ if item i requires I, for k=2 or 4,}$$

$$Pr(X_{ni} = 1 / \theta_n, \delta, \phi_{n2} = 1) = \frac{exp(\theta_n - \delta_i - \tau_R)}{1 + exp(\theta_n - \delta_i - \tau_R)}, \text{if item i requires R, for k=3 or 4,} \quad (10)$$

$$Pr(X_{ni} = 1 / \theta_n, \delta, \phi_{n2} = 1) = \frac{exp(\theta_n - \delta_i - \tau_{RI})}{1 + exp(\theta_n - \delta_i - \tau_{RI})}., \text{ if item I requires both I and R,}$$

$$Pr(X_{ni} = 1 / \theta_n, \delta, \phi_{n1} = 1) = \frac{exp(\theta_n - \delta_i)}{1 + exp(\theta_n - \delta_i)}, \text{ otherwise.}$$

This is a general measurement formulation that one might apply under a wide range of contexts when R and I were being learned and/or experimentally manipulated. For example, use of R and I might be conceptualized as "natural" (i.e., not manipulated), in which case the classes 1 through 4 are latent, and membership in them must be estimated. In the specific situation described above, the classes were experimentally determined, hence coulod be considered as known, not estimated. One way to conceptualize how the observed differences in learning would be manifest in the results from the measurement model would be to estimate the model separately in the two treatment classes training for I first as opposed to treatment for R first. Then the observation that transition from [I+ R-] to [I+ R+] is easier than the transition from [I- R+] would be captured in a finding that τ_{RI} is smaller in the I first treatment class than in the R first treatment class. This finding can be tested for statistical significance using the standard errors that are generated as part of the analysis. It can also be assessed for substantive significance (i.e., effect size) by using

the criterion-referencing techniques used in Rasch scaling, and also by combining it with the other (knowledge access and item difficulty) parameters to display the effect upon selected individuals (in terms of odds ratios etc.).

A More General Measurement Approach

We now give a more general and formal characterization of measurement from the knowledge level. We will assume that access to knowledge, within a particular knowledge-state, varies continuously, but perhaps *multidimensionally*. That is, within a particular knowledge-content state, k, at the knowledge level we may represent an agent by a D-dimensional vector of (possibly unknown) agent-parameters, $\underset{\sim}{\theta}$, and we will call these variables *components*,

$$\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_N). \tag{11}$$

For instance, for a group of people who know a particular problem-solving strategy, or who know a specific set of instructions, we may array them along a continuous scale to represent their proficiency in accessing and using that knowledge. For instance, we might parameterize the initial performance of each of N knowledge elements of cognitive skill for Lisp programming using a vector of continuous variables $\underset{\sim}{\theta}$. As indicated above, in Equations 5 and 7, we assume a a K-dimensional vector, $\underset{\sim}{\phi}$, representing K knowledge states, and a K-dimensional vector, $\underset{\sim}{\pi}$, representing the probabilities of being in the K states. Also, suppose that the environment in which these variables operate can be represented by a P-dimensional vector of (possibly unknown) environment parameters, $\underset{\sim}{\xi}$; which are traditionally termed "item parameters":,

$$\underset{\sim}{\xi} = (\xi_1, \xi_2, \ldots, \xi_P). \tag{12}$$

The examples above have both been restricted to dichotomous responses at the item level. This is not required, although it does make for less complicated sets of parameters, and consequently, for simpler interpretations. Generally, the approach we describe would be amenable to application with any of the wide variaety of models generally termed item response models. For the particular examples shown below, and in order to preserve the interpretability described above as pertaining to Rasch models, we will use a class of generalized polytomous Rasch models. Specifically, we will use a form of the Rasch model that uses a linear model on the environment parameters which includes a wide class of existing Rasch models: the *multidimensional random coefficients multinomial logit* (MRCML) model, which builds on prior work (Glas, 1990; Wang & Wilson, 1993; Wilson & Adams, in press) that permits the generation and fitting of many models in the Rasch family. As this particular model is not the focus of this paper (it

could be replaced by other formulations) we will not describe it here, but refer the reader to the Appendix for an account of the model. In general, the knowledge-level observer will be interested in responses that occur in some specific set of environmental situations which will be indexed by $i = 1, 2, \ldots I$. For now, we will assume that there is a known function f_{ki} such that in each situation i, the probability of a particular response by an agent can be represented as the probability of a realization x_i of a random variable X_i: $f_{ki}(x_i;\underset{\sim}{\xi}|\underset{\sim}{\theta})$. Note that the probability depends upon the fixed (but unknown) values of the environment ingredients $\underset{\sim}{\xi}$ and on the values of the components from the random variable $\underset{\sim}{\theta}$.

When $f_{ki}(x_i;\underset{\sim}{\xi}|\underset{\sim}{\theta})$ is a continuous function of its arguments, we can see this formulation as representing degree of knowledge access within a particular knowledge-content class. For instance, below we will discuss an example involving the learning of cognitive skills for Lisp programming. Within that knowledge-content class containing the set of cognitive skills relevant to Lisp, we will be concerned with measurements indicating improvements in access to those skills due to environmental elements, such as available examples, and amount of exposure to practice. Such knowledge-access improvements are what we associate with symbol-level learning.[10]

For instance, in the illustrative example surrounding Equations 1 to 4, we assumed the measurement of difficulties, δ_i, of exhibiting cognitive skill in appropriate situations, and these difficulties may comprise a subset of the $\underset{\sim}{\xi}$ parameters. In that example, there was only one knowledge-content class, $\phi_n = (1)$, one environment parameter, $\underset{\sim}{\xi} = (\delta_i)$, and the person knowledge access parameter was one-dimensional, $\underset{\sim}{\theta} = (\theta_n)$. In our first model application below, in which we address the odds of observing the successful (or unsuccessful) execution of a cognitive skill across learning opportunity trials, we measure learning difficulties in addition to such difficulty parameters. In the Kessler and Anderson example (Equation 10), the person knowledge access parameter is still one-dimensional, $\underset{\sim}{\theta} = (\theta_n)$. But the knowledge content class has expanded to have four elements: $\phi_n = (\phi_{n1}, \phi_{n2}, \phi_{n3}, \phi_{n4})$, and the environment parameters have been expanded to include effect-parameters, τ, as well as a set of item parameters, δ_1, δ_2, ...δ_K so that $\xi = (\delta_1, \delta_2, \ldots \delta_K, \tau_I, \tau_R, \tau_{IR})$. In real applications one would expect to have many more types of item parameters, of course.

---

[10]It is perhaps worth noting that according to our definitions, the measured improvements in knowledge access may be determined by either situational variables (e.g., improved examples for Lisp) or agent variables (e.g., greater amounts of practice). Convention in psychology restricts "learning" to organisms, but the notion of environmental learning can be found in other fields such as organizational learning.

We might characterize the effect of knowledge-level learning in the following way. Suppose that, in the ideal, if an agent belonged to just one knowledge-content state, k, of the K knowledge classes, or $\phi_k = 1$, then we could characterize the probability of response x for that agent in that context as:

$$\Pr(X_i = x | \underset{\sim}{\phi}, \underset{\sim}{\theta}, \underset{\sim}{\xi}) = \prod_{k=1}^{K} \left[f_{ki}(x; \underset{\sim}{\xi} | \underset{\sim}{\theta}) \right]^{\phi_k} \tag{13}$$

where the conditional probability f subscripted by k will differ from one knowledge-content class to another, in addition to the situation index i. Note that when the knowledge content class is known, the use of the iterative product in Equation 13 is a convenience: The product of the k - 1 terms coded with $\phi_k = 0$ will be one and the conditional probability will be simply to the term coded with $\phi_k = 1$.

Note that, although the formal assumptions of the model require that each agent belong to just one knowledge class, the results of an estimation (formally, the *posterior* distribution) will be expressed in terms of the probability of each agent being in each of the classes, which will generally *not* indicate exclusive membership in just one class. This is typical of latent class formulations. Thus, this formulation should not be seen as being inconsistent with approaches that assume that agents belong to more than one knowledge-content class at a time, such as might occur in models that assume that a person might use different strategies in the same kind of situation.

The formulation in Equation 13 is referred to as a *mixture distribution* (Titterington, et al., 1985), and techniques are available for parameter estimation when certain conditions are met (Mislevy & Verhelst, 1990). First, one must specify the functional form of the conditional probabilities $f_{ki}(x; \underset{\sim}{\xi} | \underset{\sim}{\theta})$. Second, there needs to be substantive theory which identifies and associates environment ingredients with the pattern of responses for agents in each knowledge-content class. To restate these prerequisites more generally, we need to specify how component parameters of the individual and environment parameters are mapped into specific possible situations and actions.

The MCRML model needs to be further elaborated to deal with agents possibly belonging to different latent knowledge-content states, as in Equation 13. Hence we generalize the MRCML to a Mixture MRCML (which we abbreviate M^{2}RCML). The Appendix describes how *design* and *scoring* matrices must be specified for each of the k = 1, 2, ...K knowledge-content classes to map the $\underset{\sim}{\phi}$ knowledge states, $\underset{\sim}{\theta}$ person components, and $\underset{\sim}{\xi}$ environment ingredients onto responses

In the examples, we explore ways that the M^2RCML and its submodels can be applied to knowledge measurement situations of recent interest. In the first example, we discuss application of a model to capture the degree to which subjects have access to cognitive skills for programming, where skills are represented as production rules, and to capture the effects of experimental manipulations, individual differences, and skill difficulties. This application can be viewed as an example in which each subject belongs to a single knowledge-content class which is arrived at by experimental training, and we are interested in estimates of differences in knowledge access among groups with different training. In the second example, we discuss the application of a model to data on stage-like development in a Piagetian task . In this case, both knowledge-content differences in strategies and knowledge access are estimated and a notion of ordered development through stages is also captured.

Model Application 1: Training and Practice Effects on Cognitive Skills

Our first application is based on a production system model of the acquisition and transfer of Lisp programming skill (see also, Anderson, et al., 1989; Pirolli, 1991; Pirolli & Recker, 1994) An extended discussion of the development and details of this specific model application is presented in Draney, Pirolli, and Wilson (1995).

Overview of the Experiments

As mentioned above, the data for this example come from four experiments (Bielaczyc, et al., 1995; Pirolli & Recker, 1994; Recker & Pirolli, 1994) that used the CMU Lisp Tutor (Anderson, et al., 1990). In all four experiments, the analyses concentrated on data from a Lisp Tutor lesson on recursive functions. This lesson was taught to subjects after several hours of preliminary Lisp Tutor work on other programming basics. Each lesson, including the recursion lesson, required subjects to read a text chapter on a topic and then solve related code-writing problems. The lesson on recursion in the Lisp Tutor contained 10 or 12 program-coding exercises, depending on the study, and took about two to four hours for subjects to complete.

For instance, in Pirolli and Recker (1994), for each lesson with the Lisp Tutor, subjects had to read a lesson booklet, and then had to solve a set of exercise problems. Before the recursion lesson, subjects worked through six lessons covering elementary

Lisp functions, user-defined functions, predicates and conditionals, the use of user-defined sub-functions, input-output, and iteration on numeric inputs. The lesson on elementary recursive functions contained excercise problems requiring subjects to write recursive functions that operate on numeric and list inputs. The booklets for these lessons were early drafts of chapters in the textbook by Anderson, Corbett, and Reiser (1987).

Subjects would be presented with a specific programming problem, such as the task of writing the Fact function. The Lisp Tutor monitored what the student did while writing program code and compared this behavior to its store of correct and incorrect solution steps represented in production system models. On each cycle of interaction the Lisp Tutor ran its internal production system models which would determine the next programming goal. The student would enter some small portion of code (usually corresponding to the next word-like element) much as they would enter text into a word processor. The Lisp Tutor would match the student-entered code to the feasible set of correct and incorrect steps for the programming goal (represented by production rules). Frequently, more than one correct step and more than one incorrect step was possible at each goal-point. If the student code was correct, then the Lisp Tutor set a new programming goal internally and a new cycle began. If the code was incorrect, then the tutor provided feedback and reset the same goal for the next cycle. After three strikes at the same goal, the Lisp Tutor explained the appropriate step and coded it for the subject. So long as the subject wrote correct code, the Lisp Tutor remained in the background as it performed its internal categorization of inputs and setting of internal goals.

Experiment 1 of Pirolli and Recker (1994) investigated the impact of instructional examples on cognitive skill acquisition, and the transfer of practice across programming problems. The texts introducing recursion all contained an example of a recursive function in Lisp. For about half the subjects, the example recursed on an integer input (the numeric example), and for the remaining subjects the example function recursed on a list input (the list example). The numeric example was the *Sumall* function and the list example was the *Carlist* example, both in Table 1. For each example, one could identify the ideal model production rules that would be evoked to produce the solution. Assuming that subjects used the example material to provide analogies for program-writing, we expected to observe improved acquisition of the productions associated with the examples (Pirolli, 1991).

The overall number of errors in writing a program was found to be well-predicted by summing the error rates for all the productions evoked on a solution. That is, each execution of each cognitive skill represented by a production rule could be treated as an

independent event with a probability of error dependent on the particular production. The number of errors on the entire programming solution was just the sum of the error probabilities for all the production executions involved in the solution. These error probabilities associated with each production rules were found to decrease with transfer from the example material as well as with transfer from prior practice. We specify the form of these transfer functions below.

The remaining three experiments that contributed to our dataset used the same instructional materials and example conditions. Experiment 2 of Pirolli and Recker (1994) used the same materials and example conditions, but additionally collected verbal protocols, which were analyzed to investigate the correlation of *self-explanation* and *self-regulation* learning strategies with improved cognitive skill acquisition. Bielaczyc et al. (1995) used the same materials and example conditions, but split subjects into a control group replicating Pirolli and Recker (1994) Experiment 1 and a trained group that received instruction on the effective self-explanation and self-regulation strategies found in Pirolli and Recker (1994) Experiment 2. Recker and Pirolli (1994) again used the same materials and example conditions, but split subjects into a control group who read the standard materials on a computer and a group that read a hypertext version of the materials on a computer. Across all four experiments, we have two consistent example conditions and a consistent problem set that provide a set of environments for measuring knowledge acquisition and use.

Basic Model for Lisp Learning

We developed a measurement model that contained a simple scalar ability parameter, θ, to measure the individuals' propensity for learning recursive programming. A vector, $\underset{\sim}{\xi}$, of additional parameters was used to measure (a) the difficulty of specific cognitive skills, (b) the learning rate as a function of practice, and (c) the effects of example-based learning on the initial acquisition of specific cognitive skills. Participants in our studies had experience with one or the other of the recursion examples in Table 1. Their knowledge-states could be coded as,

$$\underset{\sim}{\phi} = \begin{cases} (1, 0), \text{ if Example 1 was read} \\ (0, 1), \text{ if Example 2 was read.} \end{cases} \tag{14}$$

The histories of the individual elements of cognitive skill, each represented as a production rule, i, can be traced across problem sets regardless of problem ordering. Each individual production can be associated with a sequential history of opportunities, or

trials, t, to acquire or perform a cognitive skill. So, we can index the programming situations according to both the production i appropriate for the situation, and t, an index of the opportunity number for production i. For each situation there are $H_{it} = 2$ responses: $X_{it} = 1$ if the subject performs an action consistent with production i (a "correct" action), or $X_{it} = 0$ if the subject fails to perform the action (an "error"). We will model the log odds of error for each production i.

We predict that error odds improve as a power function of trials of practice (cf. Pirolli, 1991), so the learning curves are predicted to be linear in log(Error Odds) × log(Trials) coordinates. Differences in the difficulty, δ_i, of initially acquiring specific production rules will be reflected by translations of the performance intercept of the learning curves. Our model assumes performance improves as a power function of trials. We may characterize such power-function improvements as $\exp(\Delta_t)$ over learning opportunities, t, where

$$\Delta_t = -\alpha \log(t), \tag{15}$$

and α is an estimated parameter between zero and one. We assume α is common across productions and unaffected by production difficulty or example experience (Pirolli, 1991). Further, we expect that there will be another improvement in student performance if a relevant instructional example is present in the environment (Pirolli, 1991). With two examples, the numeric example (Example 1) and the list example (Example 2) in Table 1, there will be two corresponding possible improvements : τ_1 and τ_2. One example will improve one subset of cognitive skills, the other example will improve another subset, and some cognitive skills will not be improved by either example. In summary, the parameters characterizing the Lisp programming environment can be specified as

$$\underset{\sim}{\xi} = (\delta_1, \delta_2, \ldots, \delta_I, \tau_1, \tau_2, \alpha). \tag{16}$$

The log error odds, given subject ability and environment ingredients, will be a function like

$$\begin{aligned} &\log\left[\frac{\Pr(X_{it} = 0 | \theta, \underset{\sim}{\xi}, \phi_k = 1)}{\Pr(X_{it} = 1 | \theta, \underset{\sim}{\xi}, \phi_k = 1)}\right] \\ &= \begin{cases} \theta + \delta_i - \tau_k - \alpha \log(t), & \text{if Example } k \text{ is relevant to production } i \\ \theta + \delta_i - \alpha \log(t), & \text{otherwise.} \end{cases} \end{aligned} \tag{17}$$

Thus, when a person has experienced Example k and it is relevant to production i,

$$\frac{\Pr(X_{it}=0|\theta,\underset{\sim}{\xi},\phi_k=1)}{\Pr(X_{it}=1|\theta,\underset{\sim}{\xi},\phi_k=1)}=\exp(\theta+\delta_i-\tau_k-\alpha\log(t)). \tag{18}$$

which achieves the desirable additive structure discussed in the text surrounding Equations 1 to 4. The response probabilities can be written as

$$\Pr(X_{it}=0|\theta,\underset{\sim}{\xi},\phi_k=1)=\frac{\exp[\theta+\delta_i-\tau_k-\alpha\log(t)]}{1+\exp[\theta+\delta_i-\tau_k-\alpha\log(t)]}, \tag{19}$$

and

$$\Pr(X_{it}=1|\theta,\underset{\sim}{\xi},\phi_k=1)=\frac{1}{1+\exp[\theta+\delta_i-\tau_k-\alpha\log(t)]}. \tag{20}$$

When a person has experienced Example k and it is not relevant to production i, then we simply rewrite Equatiions 18, 19, and 20 without τ_k.

Results

The Lisp Tutor matches its ideal model against student behavior at the level of each word-like atomic symbol typed by subjects. At this grain of analysis, according to the Lisp Tutor's production system model, there are approximately 50 to 800 potential states in each of the recursion problems (covering all possible paths students are expected to take). There are about 100 productions modeling the relevant target skills and possible erroneous variants on the skills. We selected 33 productions for analysis across the N = 76 participants in the four experiments. These productions represented new knowledge presented by the text on recursion or by the instructional examples given to subjects. See the Appendix for details on how Equations 19 and 20 can be expressed as special cases of Equation A1.

A fit of the above model was performed using MRCML estimation software (Adams & Wilson, in press). The relationship among the difficulties of the productions, the size of the practice effects, the ability distributions for each example group, and the size of the example effects for this model is illustrated graphically in Figure 5, all on a common logit scale. Figure 5 illustrates the relative positions of the production rules on the logit scale, with the most difficult productions, the ones on which the most errors were made, at the top of the scale, and the easier productions near the bottom. This clearly demonstrates that different production rules have distinctly different difficulties. Distributions of initial ability for the two disjoint groups receiving different examples are shown on the left-hand side of the page, with persons who saw the number recursion

example represented by a 1, and persons who saw the list recursion example represented by a 2. This demonstrates that there were notable differences in individual ability. Each person has a greater than 0.5 probability of making an error on a production whose difficulty is above their ability level, and a less than 0.5 probability of making an error on a production whose difficult is below their ability level on the page. The two groups (created by random assignment of subjects) do not differ greatly in their initial abilities.

The shift in ability for each group when that group encounters a production rule related to the example they saw is represented as a vertical line in the box at the bottom left corner of the page. The length of each line represents the size of the "boost," τ, in performance on specific productions that is due to the availability of a relevant example. To see the effect of the "boost," this line should be used as a ruler moving it up and down the logit scale as appropriate. The effects of practice for repeated trials of a production rule are also shown in the top portion of the box. Each ruler represents the effect of the number of units of practice shown below the ruler. In other words, on the second trial, since a person has had one unit of practice, that person would receive an ability boost equivalent to the length of ruler number 1. By positioning the appropriate rulers at a given person's location on the scale, it is possible to determine the effects of practice and example on the error probability for that person on productions of that type.

Insert Figure 5 about here

Figure 5 allows a number of interpretations. First, most subject ability levels are higher on the page than the majority of production rule difficulty levels. Thus, most subjects are predicted to have a less than 0.5 probability of making an error on a production rule, even the first time they attempt it. This is an accurate reflection of the observed data, in which, for the group of subjects who saw the numerical example, the proportion of error on the first trial was less than or equal to 0.5 for all but 5 production rules, and for the group who saw the list example, the proportion was less than or equal to 0.5 for all but 4 of the production rules. Second, the sizes of the practice effects indicate that error rates for most subjects on most productions tend to drop off fairly rapidly, which is also consistent with the data.

For this model, both of the example effects are statistically significant at the .05 level, as is the learning rate parameter for trials (in logarithmic units). The effect parameter for the number example is -0.37 logits, and for the list example is -0.24 logits. These

example effects are similar in magnitude, although the effect of the number example is slightly larger than the effect of the list example. Table 2 presents the predicted and observed estimates for error rates across the two example groups broken down by the type of production. Table 2 demonstrates a rather close match between the observed and predicted values, and the pattern of predictions is reflected in the observed results. The learning rate coefficient for the natural log of trial number is -0.83. The effect of the first unit of practice (i.e. the drop in production difficulty experienced on the second trial of the production) is -.83 x log(2) = -0.58. Thus, the effect of the first unit of practice on the probability of making an error was somewhat greater than the effect of seeing an example; however, the effect of practice drops off fairly rapidly as the number of trials of a production rule increased.

Insert Table 2 about here

We can check the fit both of the overall model and of individual items and persons to the model. Figure 6 illustrates the observed and expected error rates for both groups. Figure 6 is constructed by considering the sequence of goals set by the Lisp Tutor for each programming problem across the full sequence of problems. Each goal is a production trial (an opportunity to perform a production). So Trial 1 is the essentially the production trial for the production associated with the first piece of written code for the first program, Trial 2 the second piece of code, and so on. Different productions are evoked on different trials. Each observed data point in Figure 6 averages over all subjects on the production appropriate for the trial. The observed and expected curves match with an R^2 of 0.74, which is comparable to the fits of Corbett, Anderson and O'Brien of their master learning model (Corbett, et al., 1995). Specifically, Corbett, Anderson and O'Brien (1995, p. 25) obtained an R^2 of 0.72 using approximately four times as many parameters. The Corbett et al. data come from Lisp lessons that come before recursion. Performance on the Lisp Tutor recursion lessons is typically more variable than the earlier lessons. The use of Rasch family models allows the reliability of the person separation to be calculated. This is defined in the same way as the familiar Cronbach's alpha reliability which one sees in classical testing situations. In our example, for all trials of all productions combined, the person separation reliability is 0.83, which is comparable to that for standardized achievement tests. Mean square fit statistics and *t*-value fit statistics (Wright & Masters, 1982) can be calculated for each trial of each production rule, and for each person. If the model fits the data well, the *t* statistics should be approximately normally distributed with a mean of zero and a variance of approximately one for both persons and items. Mean

square fit statistics have an expected value of one for both persons and items. These fit statistics can also be used to detect individual trials or persons for which the fit of the model is especially poor. These trials or persons can then be examined in detail, to see if the cause for the lack of fit can be determined. A "rule of thumb" which is often used is that the *t* value fit statistics should lie between 2.00 and -2.00 if the person or trial fits adequately. The average value of the mean square fit statistics for production trials in our example was 1.02, and for persons the average was 0.97. In both cases, the mean was very near its expected value of one, suggesting that the model fits the data reasonably well. For the *t* values, the mean for persons was -0.07, and the standard deviation was 0.84, which are still reasonably near the theoretical expected values. However, the low value for the standard deviation suggests that there may actually be slightly less misfit than might be expected. For trials, however, the mean of the t values is .93, with a standard deviation of 0.72. The mean in this case is higher than would be expected if the model fit the data as well as we would like.

Insert Figure 6 about here

Summary

The first example illustrates several points about the measurement approach. The Lisp Tutor ITS instantiates (quite mechanically) the operation of a knowledge-level observer. It is an automatic knowledge-ascribing instrument that scored subjects according to whether or not they were behaving as if they were exhibiting particular knowledge elements. These elements constituted a single knowledge-content class. In the next example we will deal with a situation in which more than one knowledge-content class is involved. The measurement model captured separate quantities representing an individual's propensity to learn Lisp vs the difficulty of specific programming situations. The model addressed changes in knowledge access as a function of environmental factors resulting from changes in the examples available in the environment. The model also addressed changes in knowledge access due to the experiential history of the person.

Model Application 2: Development of Strategy Mixtures for Proportional Reasoning on the Balance Beam

Our second example situation comes from the literature on the development of strategies for reasoning on balance-beam problems such as those in Figure 7. Siegler (1981) proposed that childrens' solutions to such problems are manifestations of stage-like changes in reasoning strategy. Siegler postulated a series of rules to describe development on such balance beam tasks. Rule I is: Choose the side with greater weight--if weights are equal, choose neither. Rule II is: Same as I except that if weights are equal, choose the side with greater distance--if distances also equal, choose neither. Rule III is the same as for II, but if neither the weights nor distances are equal, muddle through. Rule IV is the same as using the correct formula.

Insert Figure 7 about here

More recently, the general argument (Siegler, 1994) has been that children exhibit multiple strategies whose propensities change with experience: "cognitive change is better thought of in terms of changing distributions of ways of thinking than in terms of sudden shifts from one way of thinking to another" (Siegler, 1994, p. 2). This is entirely in line with sort of probabilistic approach that we are suggesting in this paper. Indeed, the notion that variability and adaptive change go hand in hand is the foundation of modern evolutionary explanations of behavior (Smith & Winterhalder, 1992a; Stephens & Krebs, 1986). These ideas suggest that we should be interested in modeling the changing distributions of knowledge elements and their use as agents adapt to their environment, which is exactly the situation that the M^2RCML model addresses.

Wilson's (1989) Saltus model is a special case of the M^2RCML model. Saltus was developed to address group-like cognitive development. Each subject is characterized by two variables, one quantitative and the other qualitative. The quantitative parameter, θ, indicates *degree* of proficiency, while the qualitative parameter, ϕ, denoting group membership, indicates the *nature* of proficiency. The Saltus model for hierarchical development generalizes the Rasch model for dichotomous test items (Rasch, 1960/1980) by positing H developmental groups. An agent is assumed to be in exactly one group at the time of testing, but group membership is not directly observed. Problem situations are

also classified into H classes. It is assumed that a Rasch model holds within each developmental group, and the relative distances among problem situation difficulties within a given problem situation class are the same regardless of developmental group. The relative difficulties among problem situation classes may differ from one developmental group to another, however. The amounts by which difficulties of problem situation classes vary for different groups are the "Saltus parameters": τ. Saltus parameters can capture how certain types of problem situations become much easier relative to others as people add to or reconceptualize their knowledge content, or how some problem situations actually become harder as people progress from an earlier group to a more advanced one because they previously answered correctly or incorrectly for the wrong reasons.

Under Saltus, as in M^2RCML, an agent is characterized by not just a proficiency parameter θ, but also a group membership parameter $\underset{\sim}{\phi}$. As before, if there are H potential developmental groups, then $\underset{\sim}{\phi} = (\phi_1, \ldots, \phi_H)$. ϕ_h takes the value of 1 if the agent is in Stage h and 0 if not, but posteriors for ϕ do not have this constraint. As with θ, values of $\underset{\sim}{\phi}$ are not observable but are estimated from the data. Within each group, items are governed by a Rasch model--each item i has a difficulty parameter δ_i. It is also assumed that each item, based on psychological theory, can be associated with a unique developmental group.

$\underset{\sim}{T} = ((\tau_{hk}))$ is an H-by-H matrix of Saltus parameters. In particular, τ_{hk} expresses an effect upon the difficulty of items in class k that applies to agents in developmental group h. For identification purposes, we assume $\tau_{1k}=0$, and $\tau_{h1}=0$. The probability that an agent with group membership parameter $\underset{\sim}{\phi}$, with $\phi_h=1$, and proficiency θ will respond correctly to item i, known to be in class k, is given as

$$P(x_i = 1 | \theta, \underset{\sim}{\varphi}, \delta_i, \underset{\sim}{T}) = \exp(\theta - \delta_i + \tau_{hk}) / \gamma \qquad (21)$$

where γ is the appropriate norming constant. For estimation purposes, we assume a population in which the proportion of agents in each developmental group h is π_h, with $0<\pi_h<1$. See Mislevy and Wilson (1996) for a more extensive account of this model.

We present a re-analysis of a portion of Siegler's (1981) balance beam data described in Wilson (1989). In balance beam problems, various combinations of weight are placed at various distances from the central fulcrum of a balance beam and subjects are

asked which side of the beam will go down. The data consist of the responses by fifty persons (agents) whose ages ranged from 5 years to adult, recorded twice approximately 6 months apart, to twelve balance beam problems. The two sets of measurements are combined for this analysis. This resulted in a total of 100 response vectors--7 of which were either zero or perfect, and were deleted so as not to distort the model comparison. The problems are related to several types of items related to Siegler's Rule Acquisition hierarchy for children's understanding of proportional reasoning. Siegler posits a group-like discontinuity between successive levels of understanding, made manifest by patterns of response to balance-beam problems of the sort shown in Figure 7. The three types of problem that will be used for the estimation are the following:

Dominant (D) items (items 1 to 4) are arranged so that paying attention to only the dominant characteristic of the problem (weight) will result in the correct response. Children tend to succeed on such items quite early. Weight is referred to as the "dominant dimension" in these balance beam problems, and distance as the "subordinate dimension," because children typically recognize first the salience of weight (Inhelder & Piaget, 1959; Siegler, 1981).

Subordinate (S) items (items 5 to 8) have equal numbers of weights on both sides but they are further from the fulcrum on one side. A child at an earlier group of understanding would tend to predict for S tasks that the beam would balance; at a more advanced group, the correct prediction, taking the unequal distances properly into account, would be made. Because weights are equal on both sides of the beam, it is not necessary to address the nature of the interaction of weight and distance.

Conflict-Dominant (CD) items (items 9 to 12) items have unequal weights and distances on the two sides of the fulcrum and where paying attention to only the dominant characteristic of the problem (weight) will result in the correct response.

Although the results of the estimation presented below will be based on only the data for these three types of items, Siegler actually collected data on the following item types also. We will present results showing how the estimations from the first three types may be generalised to the other three.

Equal (E) items have the same number of weights on both sides at the same distances from the fulcrum. Children recognize early on that the beam will stay balanced in this situation.

Conflict-Subordinate (CS) items have unequal weights and distances on the two sides of the fulcrum and are arranged so that paying attention to only the subordinate characteristic of the problem (distance) will result in the correct response.

Conflict-Equal (CE) items have unequal weights and distances on the two sides of the fulcrum, and in which the beam balances--a counterintuitive solution to children who recognize only the salience of weight. An expert solution requires comparing torques, or products of weights and distances.

Results

The estimated proportions in each group (note, we will use the term "group" in preference to "stage" until the groups have indeed been shown to look like stages) and the item parameter estimates for each group are given in Table 3 (note that the effect of the relevant Saltus parameters have been incorporated into these item parameter estimates). The results show that there is a non-negligible proportion of students estimated to be in each. Table 4 shows the average proportion of correct responses to items of each type that would expected from an examinee at the mean of the three group groups. This makes it easier (than in Table 3) to interpret what is going on in each of these groups. Group I is a group of people who are doing very poorly on all items. Group II is a group of people who are doing very well on both the Dominant and Conflict -Dominant items, but very poorly on the Subordinate items. Group III is a group who are doing very well on both Dominant and Subordinate items, and somewhat less well on the Conflict-Dominant items.

Insert Tables 3 and 4 about here

These groups can be further interpreted by examining which persons are categorized into each group by the estimation, which is done by assigning each person into the group that had the greatest probability of the three as shown in Table 5. The minimum probablity of assignment into these groups was .80. Table 5 shows that the persons in groups 1 and 2 are all confined to the younger age groups, while group 3 is primarily composed of those in the older age groups. This information, along with the probabilities in Table 4, helps us to interpret that the persons in group 2 are those using Siegler's Rule I, while those in group 3 are those using the higher Rules. Given the probabilities, it seems most likeley that those in group I are using either none of Siegler's Rules at all, or are using one that is even earlier in a developmemntal sense.

Insert Table 5 about here

We can gain more interpretive information by looking somewhat more closely at the behavior of the members of each of these groups: We can examine their responses to individual items, as shown in Table 6. Note in this Table, that the correct response is indicated with an asterisk. Looking first at Group 2, note that these persons make the correct response very consistently for the Dominant and Conflict-Dominant items, and equally consistently make the wrong response "Same" for the Subordinate items. This indicates that these persons are indeed responding just as one would expect of a person using Seigler's Rule I. However, we can also examine their responses to the other three item types: Equal, Conflict-Equal and Conflict-Subordinate. Note that the data regarding these three item types was not used in the estimation, so that any verification of findings actually constitutes a construct validity check on the results. As one would predict for persons using Rule I, they do indeed do very poorly on both the CE and the CS items, and do very well on the E items. These results indicate that the Group 2 group do indeed correspond very closely to those who are using Siegler's Rule I.

Insert Table 6 about here

Looking now at the Group 1 persons, we see that they are doing well on only the E and CS items. For all the others, they are doing quite poorly. Apart from the success on the CS items, this would indicate that this small group (composed of 4 to 8 year olds) is operating at an even lower level than Siegler hypothesized with his Rule I--perhaps a Rule 0, where persons respond correctly only on the easiest of questions, should be considered. The success of this group on the CS items needs some attention however. This is one of the hardest item types, so it is puzzling that they should find them relatively easy. The answer may be suggested by examining their responses in a bit more detail. They are responding to the S items like a person using Rule I (i.e.,they are getting them incorrect), yet, for the other item types (apart from the E items, which they are getting correct), they are responding the *opposite* to those in Group 2 (which corresponds to them getting the items incorrect, apart from the CS items). Now, this is a bit strange, but it does seem to be a consistent observation. Perhaps this is what is going on: they are a group who for some reason are responding the opposite of what they intended, or for whom the whole concept is so fragile that they are getting it around the wrong way.

Turning to Group 3, we can see that there is a very high degree of success on the first three item types. It looks like they have certainly mastered Rule II. Just looking at the item types in the estimation data set does not allow us to distinguish whether they have gone beyond Rule II, but examination of the results for CS shows that indeed they are tending to get these right at about a chance level (about 1/3, or approximately 21, in each

response category). The same is not so chracteristic of the CS items--they are getting two of the items right at a rate that is somewhat higher than chance, and it looks like the more heavily weighted side (the left in each case) is more of a distractor here. This looks pretty much like what Siegler meant by "muddle through" for Rule III.

Table 7 illustrates the information available for inferences about individual persons from the analysis. We illustrate four cases, the first three of which were selected as typical for each successive group. Note that the probabilities of group membership across many different data sets have usually been found to be not so extreme as the ones displayed here, but are typical in this analysis. The fourth is actually the case with the lowest maximum probability. For a given response pattern, there is a posterior probability for membership in each of the classes (and we have shown just the most likely), and, conditional on membership in a class, a location and its standard error.

Insert Table 7 about here

Summary

In this second example, observations of individuals were used to assess the degree to which they belonged to different strategy-use classes. This is because different patterns of responses are predicted from membership in different knowledge-content classes. Individuals, however, are not exclusively assigned to single categories, but are viewed as behaving consistent with a probabilistic mixture of knowledge-content classes. One may view this as a manifestation of the individual being a bundle of stochastic strategies, a result of the uncertainty of our observations, or both.

General Discussion

An essential aspect of our measurement approach is the Newell-Dennett framework that defines a knowledge level of observation and explanation. This framework defines knowledge as a relation between characteristics of an agent and the characteristics of the environment. The framework also defines the role of the observer in this framework in attributing knowledge to the agent based on manifest situations and behavior. We assume that (latent) states or classes of knowledge content are discrete, and that individual agents may be viewed as belonging to (or having) different knowledge-content classes, or to mixtures of knowledge-content classes. We assume that within these knowledge-content classes individuals in particular environmental situations have varying degrees of access to that knowledge. These (latent) knowledge-access functions are viewed as continuous.

Observed behavior will be the result of the state or mixture of knowledge content that the individual is in and the degree of knowledge access. Knowledge-content classes may be defined along the lines of such traditional distinctions as those among different stages, levels of expertise, strategies, cognitive skills, subject-matter knowledge, and so on. Knowledge-access functions may characterize traditional properties such as strength or activation. Critical to our notion of *measuring* knowledge is the development of an approach in which separate empirical quantities characterizing individuals and characterizing environments can be measured from the exhibition of knowledge, even though knowledge is defined as a relation between agents and environments. This was achieved using measurement models that derive from specifically objective measurement (Rasch, 1960; Rasch, 1977).

Heuristic Power of Measurement at the Knowledge-Level

In short, our proposal is a form of "urbane verificationism" (Dennett, 1991) that treats knowledge as a determinant of response functions inferred from observed behavior. The proposal adopts an ontological and epistemological stance on knowledge, that traces through modern proponents, such as Newell (1990) and Dennett (1988) back to Brentano (1874/1973), and aims to be broadly consistent with a variety of psychological approaches to theories of cognition. The elaboration of explanations at the knowledge level with quantitative measurement aims to dispel the informality of such explanation while improving its recognized heuristic value.

Let us consider this last point in more detail. There are a number of ways to think about the role of the knowledge level in theorizing about cognition. One common view is that knowledge level explanation is just the everyday "folk psychology" by which we make predictions about the actions of others based on our attributions about their knowledge and goals. As such, it might be a good source of interesting hypotheses about human nature, but plays no formal role in scientific explanation. Another view, exhibited especially in Dennett's (1981; 1988) earlier works is that knowledge-level explanations are *instrumental* scientific theories in a number of senses. Intelligent behavior can be reliably predicted by treating agents *as if* they are knowledge-level systems. Similarly, mechanistic (symbol-level) analyses must explain how mechanisms perform *as if* they are knowledge-level systems. The utility of knowledge-level explanations, however, is not restricted to merely obliging the limited ability of psychologists to model the full complexity of the world. Dennett (1981) argues that the knowledge level provides a principled way--not available at the symbol level--for developing the effective abstractions about human psychology that, for instance, allow us to identify the type of activity

observed rather than the mere description of token physical states and movements. Moreover, knowledge level analyses cannot be applied ubiquitously to all objects with equal effectiveness and scientific meaningfulness: such explanations for the behavior of rocks or thermostats do not work as well as they do for humans and other higher organisms. According to Dennett (1981; 1988), then, knowledge level explanations are instrumental in scientifically understanding behavior, they rationalize mechanistic accounts of intelligent behavior, and they seem to work quite well.

That knowledge-level explanations should work so well begs questioning their status as "merely" instrumental. Bechtel (1985), for instance, argues that knowledge-level explanations are realist explanations, and furthermore that they fit properly in the realm of scientific evolutionary-ecological explanations--a point of view that seems to characterize Dennett's more recent thinking (Dennett, 1995). It could be argued that the knowledge level fits evolutionary-ecological explanations in several senses. Beliefs or knowledge about the environment, preferences, and principles of rationality must ultimately be explained in terms of biological and cultural evolution (see, Smith & Winterhalder, 1992b , pp. 45-50). So, evolutionary-ecological explanations tell us why the behavior of knowledge level systems is or is not adaptive, and knowledge-level explanations in turn rationalize mechanistic accounts. Furthermore, some (Dennett, 1995) have argued that the intentionality assumed in knowledge-level systems is an expected product of evolution, and more specifically a feature that evolution has dealt to humans.

So, there are at least three interpretations of why knowledge-level explanations have heuristic power in the progress of psychology. The knowledge level either (a) provides a voluminous source of hypotheses from folk theory, (b) is a scientific level of explanation in its own right, but "merely" instrumental, or (c) a natural outgrowth of evolutionary science, or even evolution itself, and a necessary part of neoDarwinian explanation. The heuristic value of the strong latter position lies both in its commitment to the knowledge level as a valid scientific level of explanation and in its ties to adaptationism (Anderson, 1990; Kitcher, 1987). We believe that our proposed elaboration of knowledge-level explanations with quantitative measurement strengthens the empirical grounds for such explanation and, moreover, increases its heuristic power in several ways.

In addition to the heuristic power of the search for empirical quantities (see the Michell, 1990, quote above), the partial separation of knowledge-level explanation from symbol-level explanation allows for more cumulative progress. Our attempts to specify the quantitative nature of *what* is observed are coupled with knowledge-level explanations

of *why* those observations arise. Such explanation is complementary to explanations at the symbol level (ie., mechanistic process models) of *how* the observations arise (Anderson, et al., 1990; Winterhalder & Smith, 1992). This suggests that the essential aspects to which we have committed do not necessarily imply a strict commitment to an information-processing approach, although we have followed that approach throughout this paper. So from one reasonable interpretation of our framework, measurements at the knowledge level are neutral with respect to the explanations specified at other levels (e.g., as to the choice of a symbolic or connectionist model to provide a mechanistic explanation at the symbol level). One consequence of this stance is that meaningful results and explanations can accumulate at the knowledge level even while controversies remain to be resolved at the mechanistic symbol level.[11]

The measurement models we used—the M^2RCML model and its many submodels—are a sufficient but not necessary aspect of our approach. Other specific estimation models might be developed to perform the same sort of parameter estimation. However, our statistical models exhibit an approach that bridges the gap between very general-purpose and group-oriented approaches, such as ANOVA, and strong theoretical models of cognition which demand individual difference parameters. The resulting measurement models are shaped directly by structural hypotheses about cognition and knowledge and hence speak directly to those hypotheses. More generally, the Rasch approach from which these specific models evolved has generated a vast array of tools and techniques that can be used to address specific methodological problems (see, for example, Fischer & Molenaar, 1995).

Assessment and Diagnosis at the Knowledge Level

In addition, our proposal could reduce the gap from scientific theory to the instruments and technology of knowledge assessment. On the one hand, much of cognitive theory over the past half-century has been concerned with the ontology, epistemology, and formalization underlying the scientific analysis of knowledge-directed behavior. On the other hand, data analysts and psychometricians have been concerned with statistical inferences of structure from responses to assessment instruments. The former camp, in which one might lump theorists as diverse as Newell (1990) or Piaget (Inhelder & Piaget, 1959) take definite ontological stances on the nature of knowledge and its acquisition, epistemological stances on how we can know the knowledge of others, and

[11]Crowther, Batchelder, and Hu (1995) provide an interesting recent example, somewhat similar in spirit to our suggestion. That work showed that a measurement-theoretic account of perceptual recognition experiments could more effectively capture all the results of a prevailing fuzzy logic model of perception (FLMP), even while the process assumptions of the FLMP could be seriously questioned.

many modern versions make specific commitments to formalization and mechanistic explanation in the form of cognitive models. The latter camp focuses instead on an empirical methodology that is intended to reveal underlying mathematical structure from observed responses. To caricature this distinction: cognitive theorists have been practically driven by the concerns of scientific explanation whereas psychometricians have been practically concerned with assessment and diagnosis. Within the psychometric camp, the Rasch approach is distinctive because of an explicit attention to measurement theory (whereas Item Response Theory applications tend to be focused on measurement technology). In particular, the notion of specifically objective measurement serves as a basis for measuring psychological variables in a meaningful way that attempts to distance their scaling from the behavior of the assessment instrument. In developing our approach, we have attempted to integrate philosophical and scientific concerns about knowledge with concerns about meaningful measurement and the assessment of individuals.

This work has direct practical consequences for cognitive diagnosis which can be seen most clearly in Example 1. The Lisp Tutor analyses of this example are quite similar to the ITS work of Corbett et al. (1995). Although it is clear that the ACT-R theory has motivated the diagnostic student modeling components of ACT tutors, many basic assumptions differ from our analyses. The running ACT tutor's student modeling modules assume a simple two-state hidden Markov model of skill acquisition in which skill elements are in either a learned or unlearned state with a simple state-transition rule, and correct and incorrect response probabilities are conditional upon state. This is dramatically different from the complexity of assumptions about knowledge-level learning and knowledge access in the ACT-R theory, which depend on such things as strength, associative activation, and cost-benefit evaluation. Thus, the learning model of the theory is entirely different and disconnected from the learning model employed in practice by the ACT ITSs.

The student models of the ACT-R ITSs diagnose individuals by employing a simple Bayesian inference scheme that updates the two-state learning model for individual productions following each student-tutor interaction. A more complex Bayesian diagnosis scheme is employed by Mislevy (1995). Inferences about the state of individual elements of knowledge are computed from observations using Bayesian inference nets (Pearl, 1988). Such Bayesian assessment technologies can be structured directly by models such as the one we developed for the Lisp Tutor studies. Indeed, Draney et al. (1995) displayed an exemplary Bayesian inference net for diagnosis based those Lisp Tutoring

models. In addition to the practical application of knowledge-based cognitive theory, one might expect the rigors of real-world problems of assessment to further drive cognitive theory.

Scope of the Measurement Approach

From the perspective of the measurement models used in the two examples, there is a clear progression of complexity. In the first example a unidimensional model was used with no latent classes. This was because we assumed that we knew the appropriate knowledge content class for each person, by knowing which experimental condition they had participated in. The measurement model captured separate quantities representing an individual's propensity to learn Lisp and the difficulty of specific programming situations. The model addressed changes in knowledge access as a function of environmental and person factors, in particular, those resulting from changes in the examples available in the environment. The model also addressed changes in knowledge access due to the experiential history of the person. What is new about this is that we have estimates of these production difficulties and individuals' Lisp learning propensities to explain using substantive theory.

In the second example, latent classes were needed (although we restricted ourselves to a single dimension), because the mapping into strategy classes (Siegler's "stages") is indirect through the students' responses to the balance beam items. Different patterns of responses are predicted from membership in different knowledge-content classes. Individuals, however, are not exclusively assigned to single categories, but are viewed as behaving consistently with a probabilistic mixture of knowledge-content classes. One may view this as a manifestation of the individual being a bundle of stochastic strategies, a result of the uncertainty of our observations, or both. What is new about this is that each individual is estimated as a particular mixture of strategies, and that is what is to be explained using substantive theory.

It would be possible to formulate multidimensional situations that involve several latent classes, although we are not sure that the substantive theory for such complexity is truly extant, so we have not proceeded to investigate such a situation at this point.

Knowledge-level Observers

To this point, we have left unexamined the issues surrounding the reliability of knowledge-level observers. Many studies that require the coding of behavior (for instance in the analysis of verbal protocols) typically report measures of intercoder reliability. The coding of behavior by an observer apparently becomes more unreliable as the descriptive

language becomes less concrete, includes the ascription of intentions, or as the cultural difference from the observed subject becomes more pronounced (Mulder & Caro, 1985). Issues such as these may raise suspicions about the degree of invariance of the frame of reference and associated measurements. One possibile approach to address these suspicions is to include the observer in the measurement theory–in a sense, to account for their knowledge-level differences. We have not done this here, but have made attempts along these lines in cases where knowledge access has been the focus, such as in Wilson and Wang (1995) and Wilson and Case (1996).

Appendix

In this appendix we give a more formal characterization of measurement model used in the analyses, which will draw upon the notation presented in Table A1, much of which was introduced in the text.

Insert Table A1 about here

We use the notion of a *scoring function* to represent the mapping of component parameters of the individual onto situation-response pairs and a ***design function*** for the mapping of ingredient parameters of the environment onto situation-response pairs. For technical purposes, we represent these functions as matrices that specify linear combinations of parameters that are associated with situation-response pairs.

We specify f_i as the Multidimensional Random Coefficients Multinomial Logit model (MRCML, Adams, Wilson, & Wang, in press) using a design matrix A and a scoring matrix B: An agent can be characterized by the situation-response probability model:

$$f_i(x_i = h; A, B, \underset{\sim}{\xi} \mid \underset{\sim}{\theta}) = \frac{\exp[(\underset{\sim}{b}'_{ih}\,\underset{\sim}{\theta} + \underset{\sim}{a}'_{ih}\,\underset{\sim}{\xi})]}{\sum_{u=1}^{M_i} \exp(\underset{\sim}{b}'_{iu}\,\underset{\sim}{\theta} + \underset{\sim}{a}'_{ih}\,\underset{\sim}{\xi})}. \qquad \text{(A1)}$$

The scoring matrix B allows the description of the score or "performance level" that is assigned to each response alternative on each of the D dimensions associated with the agent. To do this we introduce the notion of a response score b_{ihd} which gives the score level for dimension d of the observed response alternative h to situation i. The b_{ihd} can be collected in a vector as $\underset{\sim}{b}_{ih} = (b_{ih1}, b_{ih2}, \ldots, b_{ihD})'$, and the vectors can be collected into a matrix $B = (\underset{\sim}{b}'_{11}, \underset{\sim}{b}'_{12}, \ldots, \underset{\sim}{b}'_{1H_1}, \underset{\sim}{b}'_{21}, \ldots, \underset{\sim}{b}'_{2H_2}, \ldots, \underset{\sim}{b}'_{IM_I})$.

Similarly, the design matrix A is used to specify the linear combinations of the P ingredient parameters $\underset{\sim}{\xi} = (\xi_1, \xi_2, \ldots, \xi_P)'$ used in the response probability model to describe the behavior of the response categories to each situation. These linear

combinations are defined by design vectors a'_{ih}, (i=1,...,I; m=1,...,M_i) which can be denoted collectively by the design matrix $A = (a'_{11}, a'_{12}, ..., a'_{1H_1}, a'_{21}, ..., a'_{2H_2}, ..., a'_{IM_I})'$.

We now briefly describe how equations in the text can be expressed as special cases of the scoring (B) and design (A) matrices used in equation A1 above. First, consider equations 12 and 13. We have chosen a very small example, with just two productions (δ_1 and δ_2) and one example (τ_1). The required matrices are shown in Figure A1. The responses are shown under the column headed X_{it}, and these correspond to conditions of production, example, and trial, specified in the previous three columns. The index i is for productions, and the index t is for trials. Responses also correspond to whether the student has experienced a relevant example, which is indicated under the column headed k, a 0 indicating that the relevant example was not seen, and a 1 indicating that it was seen. The scoring matrix, indicated under B, is a vector in this case, because this is a unidimensional model. The parameter vector, ξ, in this case is composed of four parameters, two for production difficulty, one for the example effect, and one for the learning rate parameter α. Thus, there are four columns in the design matrix A. Whenever the response is correct (X_{it}=0), all entries are zero, which corresponds to equation 19 (i.e.., exp(0)=1). The other entries are only non-zero when the corresponding parameter should appear in equation 20. Thus, when the response is an error on production 1, a 1 appears under δ_1, and correspondingly for δ_2, otherwise it is a zero. When the example is relevant (a 1 under A), and the response is an error, no matter which production was involved, there is a 1 under τ_1, a 0 otherwise. The coefficient of α is log(t) when an error occurs, regardless of the production and example, and a 0 otherwise. Inspection of equations 19 and 20 will show that this pattern corresponds to that described in the text.

Insert Figure A1 about here

Table A2 gives the appropriate B matrix and A matrices for a small saltus example with four items, and two stages, as in equation 21. The first two items are assumed to be associated with stage 1, and the second two with stage 2. Note that $\tau_{11}=\tau_{12}=\tau_{21}=0$, as mentioned above, for identification purposes.

Insert Table A2 about here

References

Adams, R. J., Wilson, M., & Wang, W. (in press). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, ,

Agre, P. (1993). Interview with Allen Newell. *Artificial Intelligence, 59*, 415-449.

Anderson, J. R. (1983). *The architecture of cognition* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1984). Learning to program in LISP. *Cognitive Science, 8*, 87-129.

Anderson, J. R. (1989). A theory of the origins of human knowledge. *Artificial Intelligence, 40*, 313-351.

Anderson, J. R. (1990). *The adaptive character of thought* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1993). *Rules of the mind* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Boyle, C. F., Corbett, A., & Lewis, M. W. (1990). Cognitive modelling and intelligent tutoring. *Artificial Intelligence, 42*, 7-49.

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13,* 467-505.

Anderson, J. R., Corbett, A. T., & Reiser, B. J. (1987). *Essential LISP* Reading, MA: Addison-Wesley.

Anderson, J. R., Pirolli, P. L., & Farrell, R. (1988). Learning to program recursive functions. In M. Chi, R. Glaser, & M. Farr (Ed.), *The Nature of Expertise* (pp. 153-183). Hillsdale, NJ: Lawrence Erlbaum.

Andrich, D. (1988). *Rasch models for measurement* Beverly Hills, CA: Sage Publications.

Atkinson, R. C., Herrnstein, R. J., Lindzey, G., & Luce, R. D. (1988). *Stevens' handbook of experimental psychology*. New York: Wiley,

Atkinson, R. C. & Paulson, J. A. (1972). An approach to the psychology of instruction. *Psychological Bulletin, 78,* 49-61.

Bechtel, W. (1985). Realism, instrumentalism, and the intentional stance. *Cognitive Science, 9,* 473-497.

Bielaczyc, K., Pirolli, P., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13,* 221-252.

Brentano, F. (1874/1973). *Psychology from an empirical standpoint* New York: Humanities Press.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation* London: Longmans, Green, and Co.

Carroll, J. B. (1988). Individual differences in cognitive functioning. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Ed.), *Steven's handbook of experimental psychology (Vol. 2)* (pp. 813-862). Yew York: Wiley.

Chi, M. T. H., Bassok, M., & Lewis, M. W. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(145-182),

Chomsky, N. (1965). *Aspects of the theory of syntax* Cambridge, MA: MIT Press.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186-190.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modelling in the ACT Programming Tutor. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Ed.), *Cognitively diagnostic assessment* (pp. 19-41). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic models of perception. *Psychological Review, 102*, 396-408.

Dennett, D. C. (1981). True believers: The intentional strategy and why it works. In A. F. Heath (Ed.), *Scientific explanation* Oxford, England: Clarendon Press.

Dennett, D. C. (1988). *The intentional stance.* Cambridge, MA: Bradford Books, MIT Press.

Dennett, D. C. (1991). *Consciousness explained* Boston: Little, Brown, and Co.

Dennett, D. C. (1995). *Darwin's dangerous idea* New York: Simon and Schuster.

Dietterich, T. G. (1986). Learning at the knowledge level. *Machine Learning, 1*, 287-316.

Draney, K., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. Nichols, S. Chipman, & R. Brennan (Ed.), *Cognitively diagnostic assessment* (pp. 103-125). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ekstrand, K. & Wilson, M. (1990, April). *Application of the Saltus model to a Piagetian test of cognitive development.* Paper presented at the

Annual Meeting of the American Educational Research Association, Boston.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G. H. & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications* New York: Springer-Verlag.

Flavell, J. H. (1972). An analysis of cognitive-developmental sequences. *Genetic Psychology Monographs, 86,* 279-350.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* doctoral dissertation, Twente, The Netherlands: Twente University,

Glas, C. A. W. (1990). A Rasch model with multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* Norwood: NJ: Ablex.

Gustafson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *Journal of Mathematical and Statistical Psychology, 33,* 205-233.

Halliday, D. & Resnick, R. (1970). *Fundamentals of physics* New York: Wiley.

Inhelder, B. & Piaget, J. (1959). *La genese des structurs logiques elementaires: Classification et seriation [The early growth of logic in the child: Classification and seriation]* Neuchatel, France: Delachaux et Niestle.

Kelderman, H. & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kendall, M. G. & Stuart, A. (1969). *Advanced theory of statistics* London: Griffin.

Kessler, C. M. & Anderson, J. R. (1985). Learning the flow of control: recursive and iterative procedures. *Human-Computer Interaction, 2,* 135-166.

Kitcher, P. (1987). Why not the best? In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 78-102). Cambridge, MA: MIT Press.

Krantz, D. H. (1964). Conjoint measurement: The Luce-Tukey axiomatization and some extensions. *Journal of Mathematical Psychology, 1,* 248-277.

Lazersfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis* New York: Houghton-Mifflin.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores* Reading, MA: Addison-Wesley.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1,* 1-27.

Masters, G. N. & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529-544.

Michell, J. (1990). *An introduction to the logic of psychological measurement* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Ed.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. & Verhelst, N. (1990). Modelling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195-215.

Mislevy, R. J. & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of dicontinuous development. *Psychometrika, 61,* 41-71.

Mislevy, R. J., Wingersky, M. S., Irvine, S. H., & Dann, P. L. (1991). Resolving mixtures of strategies in spatial visualization tasks. *British Journal of Mathematical and Statistical Psychology, 44,* 265-288.

Moore, J. & Newell, A. (1973). How can MERLIN understand? In L. Gregg (Ed.), *Knowledge and cognition* (pp. 201-252). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mulder, M. B. & Caro, T. M. (1985). The use of quantitative observational techniques in anthropology. *26*, 323-330.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18,* 87-127.

Newell, A. (1990). *Unified theories of cognition* Cambridge, MA: Harvard University Press.

Newell, A. (1993). Reflections on the knowledge level. *Artificial Intelligence, 59,* 31-38.

Newell, A., Yost, G., Laird, J. E., Rosenbloom, P. S., & Altmann, E. (1992). Formulating the problem-space computational model. In R. F. Rashid (Ed.), *CMU Computer Science: A 25th anniversary commerative* (pp. 255-293). New York: ACM Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference* Los Altos, CA: Morgan Kaufman.

Pirolli, P. (1985, *A cognitive model and intelligent computer tutor for programming recursion.* Paper presented at the Cognitive Science Seminar and the SESAME Colloquium, University of California, Berkeley.

Pirolli, P. (1991). Effects of examples and their explanations in a lesson on recursion: A production system analysis. *Cognition and Instruction, 8,* 207-259.

Pirolli, P. & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12,* 235-275.

Pirolli, P. L. & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39,* 240-272.

Polson, P. G., Bovair, S., & Kieras, D. (1987). Transfer between text editors. In J. M. Carroll & P. Tanner (Ed.), Proceedings of the *Proceddings of CHI '87 Human Factors in Computing Systems and Graphics Interface Conference* (pp. New York.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14,* 58-94.

Recker, M. & Pirolli, P. (1994). Modelling individual differences in students' learning strategies. *Journal of the Learning Sciences, 4,* 1-38.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monograph of the Society for Research in Child Develpment, 46,*

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science, 3,* 1-5.

Singley, M. K. & Anderson, J. R. (1989). *Transfer of cognitive skill* Cambridge, MA: Harvard University Press.

Smith, E. A. & Winterhalder, B. (1992a). *Evolutionary ecology and human behavior*. New York: de Gruyter,

Smith, E. A. & Winterhalder, B. (1992b). Natural selection and decision-making: Some fundamental principles. In E. A. Smith & B. Winterhalder (Ed.), *Evolutionary ecology and human behavior* (pp. 25-60). New York: de Gruyter.

Stephens, D. W. & Krebs, J. R. (1986). *Foraging theory* Princeton, NJ: Princeton University Press.

Stevens, S. S. (1951). *Handbook of experimental psychology* New York: Wiley.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions* Chichester, UK: John Wiley and Sons.

Wang, W. & Wilson, M. (1993, April). *Comparing multiple-choice items and performance-based items using item response modeling.* Paper presented at the Sixth International Objective Measurement Workshop, Atlanta, GA.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in development. *Psychological Bulletin, 105*(2), 276-289.

Wilson, M. (1993, June). *A multilevel perspective on quasi-experimental research.* Paper presented at the Annual Meeting of the Psychometric Society, Berkeley, CA.

Wilson, M. & Adams, R. J. (1992, July). *A multilevel perspective on the "two scientific disciplines of psychology".* Paper presented at the a symposium at the XXV International Congress of Psychology, Brussells.

Wilson, M. & Adams, R. J. (in press). Rasch models for item bundles. *Psychometrika.*

Wilson, M. & Case, H. (1996). *An examination of rater performance over time* (Tech. Rep. BEAR 96-2). Berkeley, CA: University of California, Graduate School of Education.

Wilson, M. & Draney, K. (1995, April). *Partial credit in the developmental context: A mixture model approach.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wilson, M. & Pirolli, P. (1995). *The relationship between the Rasch model and conjoint measurement structures* (Tech. Rep. . Berkeley, CA: University of California.

Wilson, M. & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessement. *Applied Psychological Measurement, 19,* 51-72.

Winterhalder, B. & Smith, E. A. (1992). Evolutionary ecology and the social sciences. In E. A. Smith & B. Winterhalder (Ed.), *Evolutionary ecology and human behavior* (pp. 3-23). New York: de Gruyer.

Wright, B. D. & Douglas, G. A. (1977a). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281-295.

Wright, B. D. & Douglas, G. A. (1977b). Conditional versus unconditional procedures for sample-free analysis. *Educational and Psychological Measurement, 37,* 573-586.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis* Chicago: MESA Press.

Author Notes

This research has been funded by the Office of Naval Research, Cognitive Science Program, grant no. N00014-91J-1523. The authors are listed in alphabetical order. We would like to thank Kate Bielaczyc and Mimi Recker for providing the data for Example 1, Robert Siegler for the data for Example 2 , and Karen Draney for performing the analyses in Examples 1 and 2.

We also thank Jim Greeno for encouraging discussions at the early phases of this work. We thank John Anderson, Jim Greeno, Robert Siegler, and an anonymous reviewer for their extensive constructive criticism.

Table 1

General patterns of transfer of knowledge from examples to Lisp programming problems .

	Number Example	List Example
	Sumall	*Carlist*
Number Problem		
Fact	High	Low
List Problem		
Length	Low	High

Table 2

Observed and predicted proportion errors per production in the Lisp Tutor studies.

	Group			
	Number Recursion Example		List Recursion Example	
Production Type	Observed	Predicted	Observed	Predicted
Number Example	.26	.27	.27	.32
List Example	.18	.21	.12	.16
Both Examples	.26	.29	.28	.30
Neither Example	.27	.25	.25	.23

Table 3

Proportions and item parameter estimates for Siegler's (1981) data

	Group		
	1	2	3
Proportion	0.08	0.28	0.64
D1	-1.18	-1.18	-1.18
D2	-0.83	-0.83	-0.83
D3	-1.18	-1.18	-1.18
D4	-0.04	-0.04	-0.04
S1	0.89	7.08	0.71
S2	0.35	6.54	0.17
S3	0.55	6.74	0.36
S4	0.35	6.54	0.17
CD1	0.72	-1.51	2.47
CD2	0.16	-2.06	1.92
CD3	-0.45	-2.68	1.30
CD4	0.65	-1.57	2.41

Table 4

Modeled proportion of correct responses at group means for Siegler's (1981) data

	Group		
	1	2	3
Proportion	0.08	0.28	0.64
D1	0.19	0.99	0.98
D2	0.14	0.98	0.98
D3	0.19	0.99	0.98
D4	0.07	0.96	0.95
S1	0.03	0.02	0.90
S2	0.05	0.03	0.94
S3	0.04	0.02	0.93
S4	0.05	0.03	0.94
CD1	0.03	0.99	0.60
CD2	0.06	0.99	0.73
CD3	0.10	1.00	0.83
CD4	0.04	0.99	0.62

Table 5

Categorization of people in by age and strategy groups estimated from the balance beam solutions. See text for details.

Age	Group 1	Group 2	Group 3
4	4	13	3
5	2	15	3
8	2	-	18
12	-	-	20
21	-	-	20

Table 6

Responses by persons in each group to each item

	Group 1			Group 2			Group 3		
Item	Left	Right	Same	Left	Right	Same	Left	Right	Same
D1	5	2*	1	0	28*	0	1	62*	1
D2	0*	7	1	27*	1	0	64*	0	0
D3	7	0*	1	0	28*	0	0	64*	0
D4	0*	8	0	27*	4	0	64*	0	0
S1	1*	0	7	0*	0	28	57*	1	6
S2	0	0*	8	1	1*	26	3	60*	1
S3	1*	0	7	0*	1	27	59*	4	1
S4	0	0*	8	3	1*	24	2	60*	2
CD1	1*	7	0	28*	0	0	38*	11	15
CD2	1*	6	1	28*	0	0	46*	7	11
CD3	7	0*	1	0	28*	0	2	54*	8
CD4	7	1*	0	1	27*	0	9	40*	15
E1	2	0	6*	0	1	27*	3	1	60*
E2	2	0	6*	0	0	28*	1	3	60*
E3	0	0	8*	0	1	27*	2	2	60*
E4	1	0	7*	3	3	22*	3	6	55*
CE1	1	7	0*	27	1	0*	23	8	33*
CE2	0	7	1*	27	1	0*	31	14	19*
CE3	0	8	0*	27	1	0*	41	5	18*
CE4	0	8	0*	26	2	0*	19	8	37*
CS1	5*	2	1	1*	27	0	22*	21	21
CS2	0	7*	1	26	2*	0	15	31*	18
CS3	0	8*	0	27	1*	0	23	24*	17
CS4	7*	0	1	2*	26	0	21*	22	21

Table 7

Posterior distributions for typical examinees under the three-group Saltus model for Siegler's (1981) data

Response Pattern	Most Probable Group	Probability	Location	Standard Error
1000 0010 0100	1	1.00	-2.62	.00
1110 0000 1111	2	1.00	2.98	.21
1111 1111 1110	3	1.00	2.91	.18
1111 0110 1111	2	0.80	2.88	.18

Table A1

Notation used in the measurement model.

$i = 1, ..., I$	Index of situations of interest in the environment (items)
M_i	Number of response alternatives to situation i
X_i	Response random variable for situation i (or vector-valued response random variable with H_i components)
D	Number of dimensions modelling the agent
$\underset{\sim}{\theta} = (\theta_1, \theta_2, ..., \theta_D)$	*Components*: D-dimensional vector of variables associated with an agent
P	Number of dimensions modelling the environment
$\underset{\sim}{\xi} = (\xi_1, \xi_2, ..., \xi_P)$	*Environment Parameters*: P-dimensional vector of variables associated with the environment
K	Number of knowledge-content categories
$\underset{\sim}{\phi} = (\phi_1, \phi_2, ..., \phi_K)$	Vector of zeros and a single one that indicates the latent knowledge-content category involved in the response
A	*Design* matrix specifying the linear combinations of environment ingredients associated with situation-response elements.
B	*Scoring* matrix specifying the level scored on the D agent dimensions by situation-response elements

Table A2

The scoring vector and design matrices for the saltus example

			A matrix									
			for ϕ_1=1					for ϕ_2=1				
Item	Response	B	δ_1	δ_2	δ_3	δ_4	τ_{22}	δ_1	δ_2	δ_3	δ_4	τ_{22}
1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	1	0	0	0	0	1	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	1	0	0	0	0	1	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	0	1	0	0	0	0	1	1

Figure Captions

Figure 1. Instructional examples and sample problem exercises for learning to program recursion in Lisp.

Figure 2. A schema representing a situation and action taken in Lisp programming.

Figure 3. Schematic representation of achieving separation of conjointly measured variables by tranformation from a non-additive structure to an additive one.

Figure 4. Transitions paths among knowledge-content states for learning iteration and recursion (see text for details).

Figure 5. Production difficulties, group abilities, and practice effects for Lisp learning.

Figure 6. Modeled and observed error rates for all subjects across all production rules.

Figure 7. Balance beam problems from Siegler (1981).

Figure A1. Scoring matrix B and design matrix A for a production system model with i= {1, 2} productions, in k = {0, 1} conditions, over t = {1, 2} trials.

Number Example

Sumall takes a positive integer, *n*, as input and computes the sum of all integers 0,..., *n*.

```
(defun sumall (n)
   (if (equal n 0) 0
      (+ n
        (summall (- n 1)))))
```

List Example

Carlist takes a Lisp list, *l*, such as ((a b) (c d) (e f)), and computes a list containing the first elements of each embedded list, e.g., (a c e).

```
(defun carlist (l)
   (if (null l) nil
      (cons (first l)
        (carlist (rest l)))))
```

Number Problem

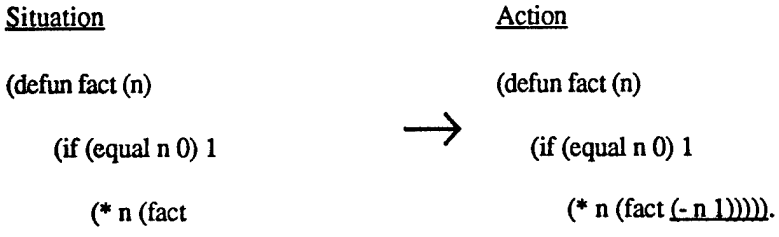
Fact takes a positive integer, *n*, and computes *n*!.

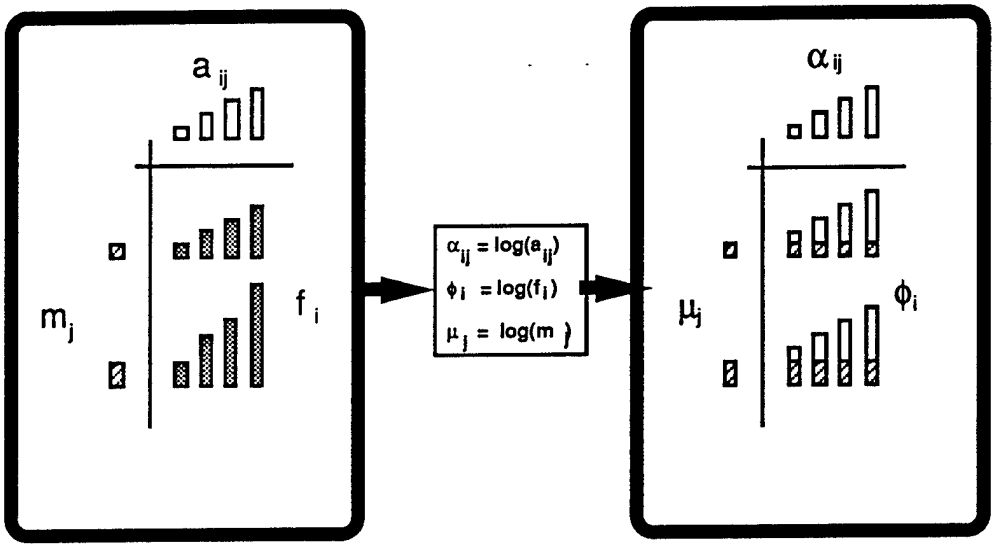
```
(defun fact (n)
   (if (equal n 0) 1
      (* n
        (fact (- n 1)))))
```

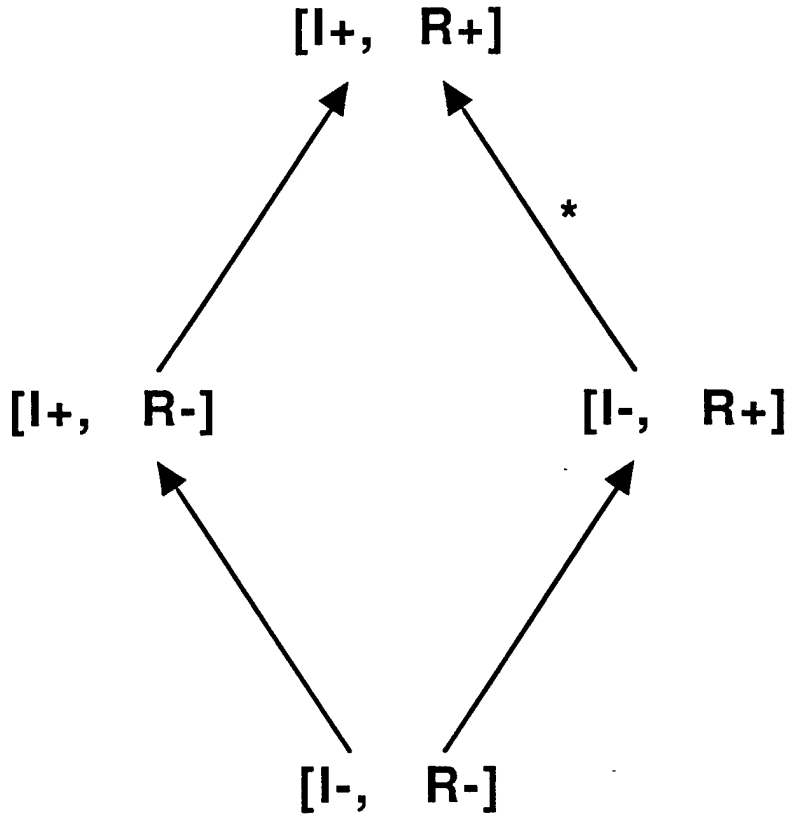
List Problem

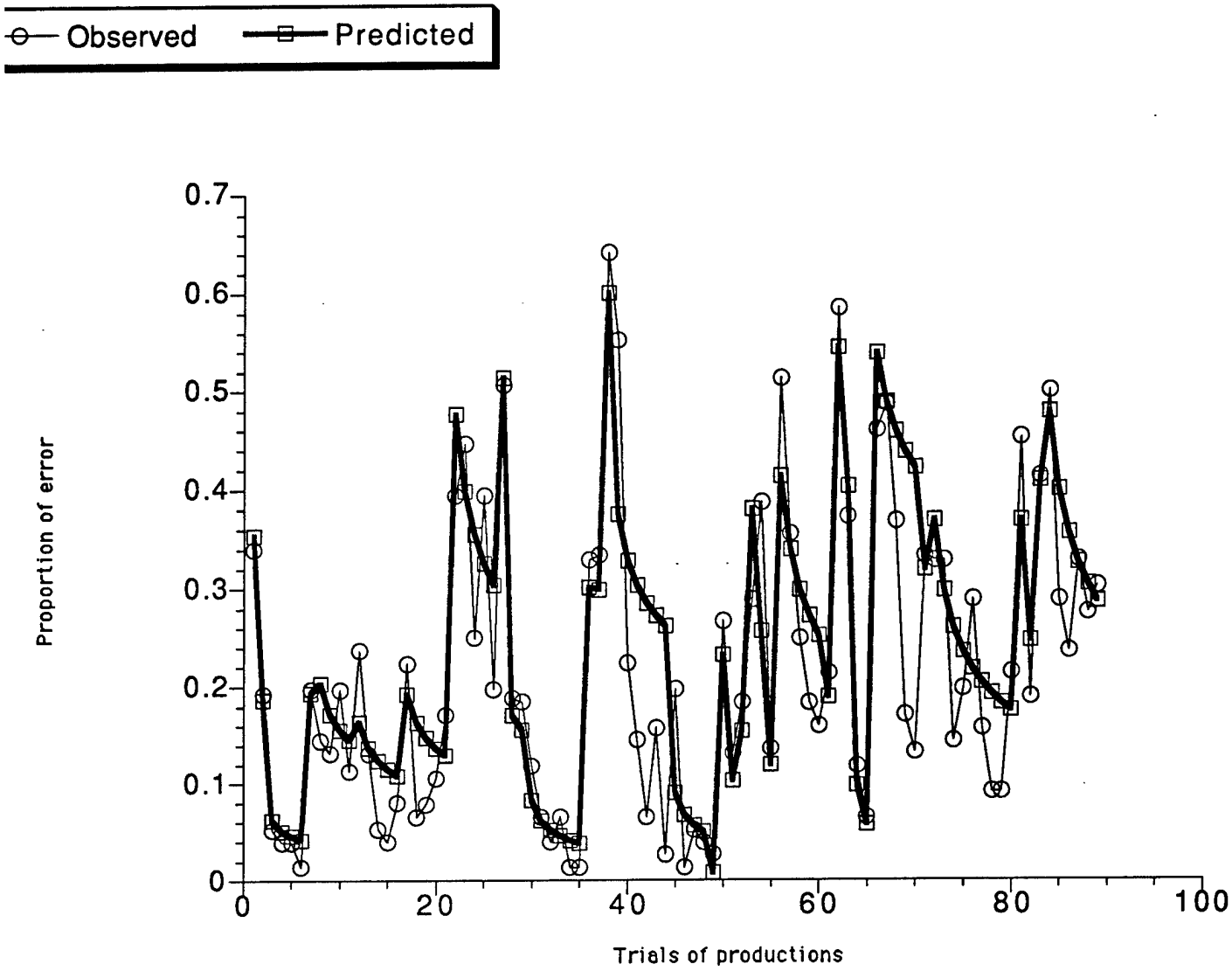
Length takes a Lisp list, *l*, as input and returns the length of the list.

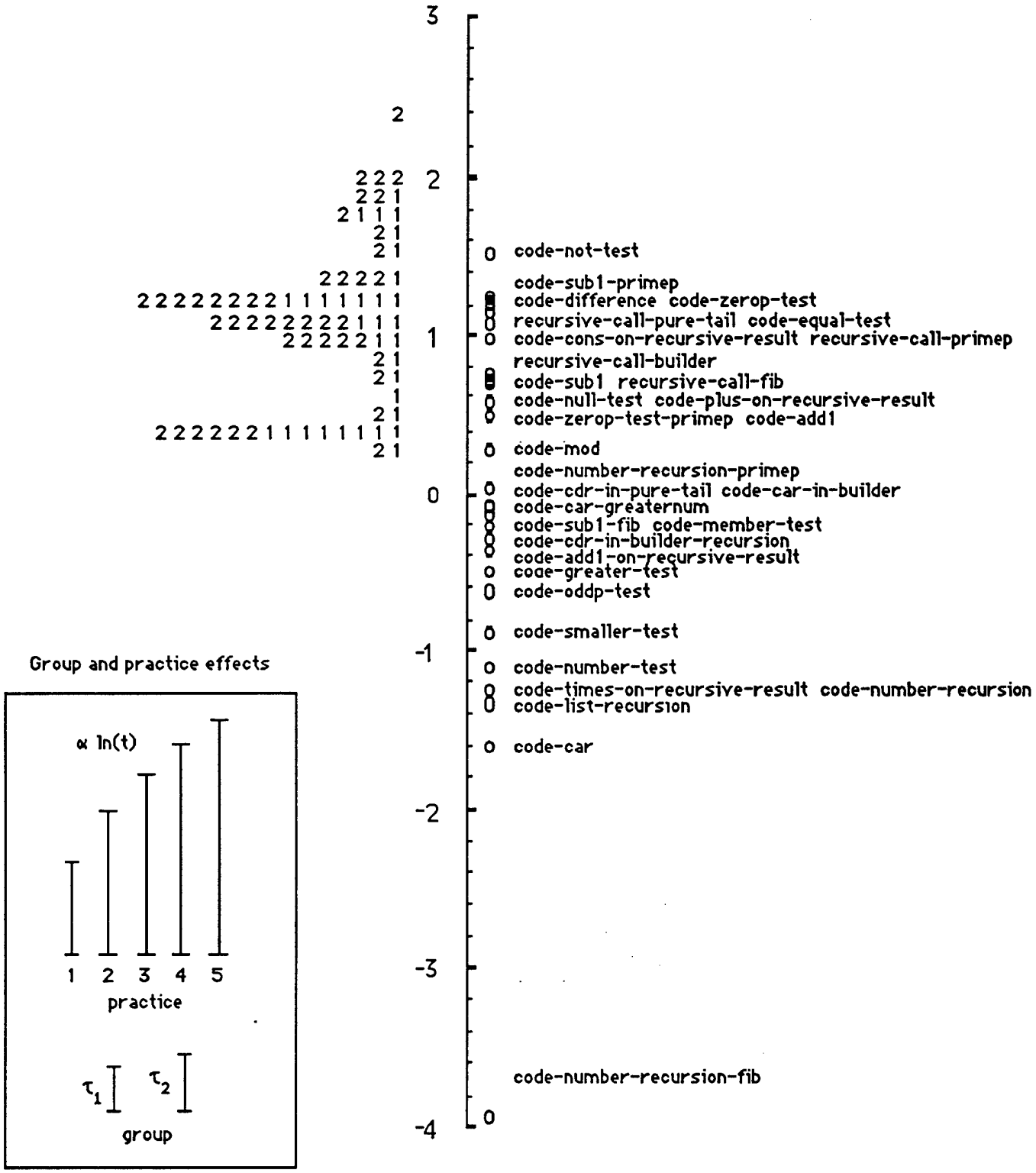
```
(defun length (l)
   (if (null l) 0
      (+ 1
        (length (rest l)))))
```


Situation
(defun fact (n)
(if (equal n 0) 1
(* n (fact
Action
(defun fact (n)
(if (equal n 0) 1
(* n (fact (- n 1)))))).

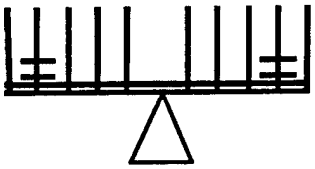

a ij
m j
f i
α ij = log(a ij)
φ i = log(f i)
μ j = log(m j)
α ij
μ j
φ i


[I+, R+]
*
[I+, R-]
[I-, R+]
[I-, R-]

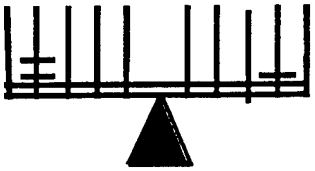

Observed
Predicted
Proportion of error
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
Trials of productions



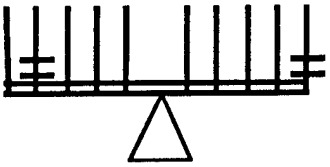
Group and practice effects



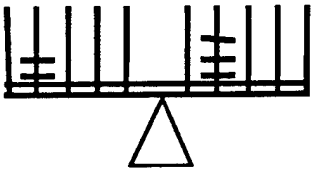
Equal



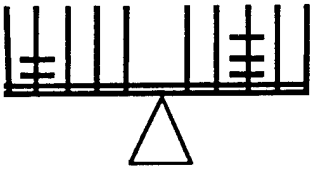
Dominant (Weight)



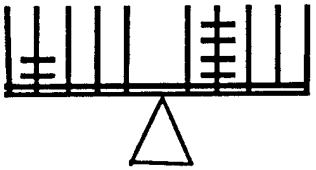
Subordinate (Length)



Conflict - Subordinate



Conflict - Dominant



Conflict - Equal

				B	A			
i	k	t	X_{it}		δ_1	δ_2	τ_1	α
1	0	1	1	1	1	0	0	log(1)
1	0	1	0	0	0	0	0	0
1	0	2	1	1	1	0	0	log(2)
1	0	2	0	0	0	0	0	0
2	0	1	1	1	0	1	0	log(1)
2	0	1	0	0	0	0	0	0
2	0	2	1	1	0	1	0	log(2)
2	0	2	0	0	0	0	0	0
1	1	1	1	1	1	0	1	log(1)
1	1	1	0	0	0	0	0	0
1	1	2	1	1	1	0	1	log(2)
1	1	2	0	0	0	0	0	0
2	1	1	1	1	0	1	1	log(1)
2	1	1	0	0	0	0	0	0
2	1	2	1	1	0	1	1	log(2)
2	1	2	0	0	0	0	0	0

REPORT DOCUMENTATION PAGE

Form Approved *OMB No. 0704-0188*

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1996	FINAL

4. TITLE AND SUBTITLE
A Theory of The Measurement of Knowledge Content, Access and Learning

5. FUNDING NUMBERS
G
N00014-91-J-1523

6. AUTHOR(S)
Peter Pirolli and Mark Wilson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Graduate School of Education
University of California, Berkeley
Berkeley, CA 94720

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Cognitive Sciences Program
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

We develop an approach to the measurement of knowledge content, knowledge access, and knowledge learning. This approach has two elements: First we describe a theoretical view of cognition, called the Newell-Dennett framework, which we see as being particularly favourable to the development of a measurement approach. Then, we describe a class of measurement models, based on Rasch modeling, which we see as being particularly favourable to the development of cognitive theories. Knowledge content and access are viewed as determining the observable actions selected by an agent in order to achieve desired goals in observable situations. To the degree that models within the theory fit the data at hand, one considers measures of observed behavior to be manifestations of intelligent agents having specific classes of knowledge content and varying degrees of access to that knowledge. Although agents, environment, and knowledge are constitutively defined (in terms of one another), successful application of our theory affords separation of parameters associated with the person from those associated with the environment. We present and discuss two examples of measurement models developed within our approach that address the evolution of cognitive skill, strategy choice and application, and developmental changes in mixtures of strategy use.

14. SUBJECT TERMS
Measurement, Knowledge Content, Knowledge Access, Learning

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	